Modern Methods of Nonlinear Optimization (Optimal Control) *

Martin Brokate **

Contents

*Lecture Notes, SS 2015

**Zentrum Mathematik, TU München

1 Introduction

Example 1.1
A car starts at point 0 from zero speed. It has to reach a point $y_1 > 0$ in minimal time T, and its speed at time T again has to be zero.

A simple mathematical model for this optimization problem is the following. Let $y(t)$ be the position of the car at time t, m the mass of the car, and u the accelerating (or braking) force. We want to

$$\text{minimize} \quad T \tag{1.1}$$

subject to the differential equation

$$m\ddot{y}(t) = u(t)\,, \quad t \in (0,T)\,, \tag{1.2}$$

the initial and terminal conditions

$$\begin{aligned} y(0) &= 0\,, \qquad \dot{y}(0) = 0\,, \\ y(T) &= y_1\,, \quad \dot{y}(T) = 0\,, \end{aligned} \tag{1.3}$$

and the constraint

$$|u(t)| \le u_{max}\,, \quad t \in (0,T)\,. \tag{1.4}$$

The function $u : [0,T] \to \mathbb{R}$ is called the **control**, the function $y : [0,T] \to \mathbb{R}$ is called the **state**.

The intuitively obvious solution is to use maximal acceleration and then maximal breaking,

$$u(t) = \begin{cases} u_{max}\,, & t < \frac{T}{2}\,, \\ -u_{max}\,, & t \ge \frac{T}{2}\,. \end{cases} \tag{1.5}$$

The time T depends on y_1 and can be computed. For $t \le 1/2$ we have

$$\dot{y}(t) = \frac{u_{max}}{m}t\,, \quad y(t) = \frac{u_{max}}{2m}t^2\,,$$

thus

$$y\big(\frac{T}{2}\big) = \frac{u_{max}}{8m}T^2\,,$$

and by symmetry we get

$$y_1 = y(T) = 2y\big(\frac{T}{2}\big) = \frac{u_{max}}{4m}T^2\,.$$

which we can solve for T. □

This optimization problem is **infinite dimensional**. Both the control u and the state y are functions on $[0,T]$ and thus a priori have infinitely many degrees of freedom, and (1.2) specifies infinitely many equality constraints. Another feature of this problem is that the optimal control is **discontinuous**, it has a jump at time $T/2$.

Example 1.2 (Optimal Heating)
We consider a body $\Omega \subset \mathbb{R}^3$ with boundary Γ. It is heated (or cooled) from the boundary with an external temperature source $u : \Gamma \to \mathbb{R}$, which thus depends on $x \in \Gamma$ but not on time. Within the body, the temperature distribution starts from some initial condition and converges to a stationary (that is, constant in time) temperature distribution $y : \Omega \to \mathbb{R}$. In a simplified model (when the heat conduction coefficient is assumed constant), y is a solution of the Laplace equation

$$-\Delta y = 0 \quad \text{in } \Omega. \tag{1.6}$$

The process at the boundary is described by

$$\partial_n y = \alpha(u - y)\,. \tag{1.7}$$

The temperature difference $u - y$ is proportional to the to the normal component of the temperature gradient; the constant α is called the heat transfer coefficient. The external source is restricted by

$$u_{min}(x) \le u(x) \le u_{max}(x)\,, \quad x \in \Gamma\,. \tag{1.8}$$

There may also be constraints on the temperature within Ω in the form

$$y_{min}(x) \le y(x) \le y_{max}(x)\,, \quad x \in \Gamma\,. \tag{1.9}$$

The goal is to achieve a desired temperature distribution $y_d : \Omega \to \mathbb{R}$ as close as possible. There is no unique formulation as an optimization problem for this. From the standpoint of mathematics, a convenient way is to choose

$$J(y, u) = \frac{1}{2}\int_\Omega (y(x) - y_d(x))^2\,dx + \frac{c}{2}\int_\Gamma u(x)^2\,dS(x) \tag{1.10}$$

as a **cost functional** to be minimized; here $c \ge 0$ is a freely chosen constant. □

Usually there is no hope to solve a problem like this in closed form, that is, to give an explicit formula for the optimal control.

2 Time Optimal Control

In this section we consider a classical time optimal control problem.

Problem 2.1
We want to

$$\text{minimize} \quad T \tag{2.1}$$

such that the solution $y : [0,T] \to \mathbb{R}^n$ of the initial value problem

$$\dot{y} = A(t)y + B(t)u\,, \quad y(0) = y_0 \tag{2.2}$$

satisfies the terminal condition

$$y(T) = y_1 \tag{2.3}$$

The control $u : [0,T] \to \mathbb{R}^m$ is assumed to be a measurable function which satisfies

$$u(t) \in \Omega\,, \quad \text{a.e. in } [0,T], \tag{2.4}$$

where

$$\Omega \subset \mathbb{R}^m \quad \text{is compact, convex, not empty.} \tag{2.5}$$

We assume that $A \in L^\infty(\mathbb{R}_+, \mathbb{R}^{(n,n)})$ and $B \in L^\infty(\mathbb{R}_+, \mathbb{R}^{(n,m)})$. □

In order that Problem 2.1 has a solution, it is necessary that there exists a control u such that the terminal condition $y(T) = y_1$ can be satisfied for some $T > 0$. This is a question of **controllability**.

Let us define the set of admissible controls

$$\begin{aligned} \mathcal{U} &= \{u : u \in L^\infty(\mathbb{R}_+, \mathbb{R}^m), u(t) \in \Omega \text{ a.e.}\} \\ \mathcal{U}_t &= \{u\big|_{[0,t]} : u \in \mathcal{U}\}\,, \quad t > 0\,. \end{aligned} \tag{2.6}$$

The sets $\mathcal{U}$ and $\mathcal{U}_t$ are convex, closed and bounded subsets of $L^\infty(\mathbb{R}_+, \mathbb{R}^m$ and $L^\infty(0,t;\mathbb{R}^m)$, respectively. This follows directly from the corresponding properties of Ω.

It is a result of the theory of ordinary differential equations that, under the assumptions of Problem 2.1, the initial value problem (2.3) has a unique solution $y : [0,T] \to \mathbb{R}^m$ for any given $u \in \mathcal{U}$, which is Lipschitz continuous. More precisely, there exists a the unique continuous function y which solves of the associated integral equation

$$y(t) = y_0 + \int_0^t A(s)y(s) + B(s)u(s)\,ds\,, \quad \text{for all } t > 0, \tag{2.7}$$

the Lipschitz continuity of y then follows from (2.7). By

$$y(t;u) \tag{2.8}$$

we denote the value of this solution at time t.

Definition 2.2 (Reachable set)
*The **reachable set** for Problem 2.1 is defined as*

$$\mathcal{R}(t) = \{y(t;u) : u \in \mathcal{U}\}\,. \tag{2.9}$$

Proposition 2.3
Let $t > 0$. Every sequence $(u_n)_{n\in\mathbb{N}}$ in $\mathcal{U}_t$ has a subsequence $(u_{n_k})_{k\in\mathbb{N}}$ which is weakly star convergent in $\mathcal{U}$, that is, there exists a $u \in \mathcal{U}_t$ such that

$$\lim_{k\to\infty} \int_0^t \langle u_{n_k}(s), v(s)\rangle \, ds = \int_0^t \langle u(s), v(s)\rangle \, ds\,, \quad \forall\, v \in L^1(0,t;\mathbb{R}^m)\,. \tag{2.10}$$

Proof: Since $\mathcal{U}_t$ is closed and bounded in $L^\infty(0,t;\mathbb{R}^m)$, this follows from a general result of functional analysis. □

It is another result of the theory of ordinary differential equations that the solution $y(t;u)$ of (2.3) can be represented in the form

$$y(t;u) = \Phi(t,0)y_0 + \int_0^t \Phi(t,s)B(s)u(s)\, ds\,, \tag{2.11}$$

where Φ is a matrix-valued function with $\Phi(t,s) \in \mathbb{R}^{(n,n)}$ arising from the homogeneous system $\dot{y} = A(t)y$; it satisfies

$$\sup_{t,s\in[0,M]} \|\Phi(t,s)\| < \infty \tag{2.12}$$

for all $M > 0$.

Proposition 2.4
The reachable set $\mathcal{R}(t)$ is a compact, convex and nonempty subset of $\mathbb{R}^n$, and the set

$$\bigcup_{0\le\tau\le t} \mathcal{R}(\tau) \tag{2.13}$$

is bounded for every $t \ge 0$.

Proof: Let $t \ge 0$. We have $\mathcal{R}(t) \neq \emptyset$, since $\mathcal{U} \neq \emptyset$. Formula (2.11) shows that $y(t;u)$ is convex w.r.t. u, that is,

$$y(t;\lambda u_1 + (1-\lambda)u_2) = \lambda y(t;u_1) + (1-\lambda)y(t;u_2)\,. \tag{2.14}$$

Since $\mathcal{U}$ is convex, this equation implies that $\mathcal{R}(t)$ is convex for every $t \ge 0$.

From (2.11) we get for $\tau \le t$ that

$$\|y(\tau;u)\| \le \|\Phi(\tau,0)y_0\| + \|u\|_\infty \int_0^\tau \|\Phi(\tau,s)B(s)\|\, ds\,. \tag{2.15}$$

Since $\mathcal{U}$ is bounded, we have

$$\sup_{u\in\mathcal{U}} \|u\|_\infty < \infty\,. \tag{2.16}$$

It now follows from (2.12) and (2.15) that $\cup_{0\le\tau\le t}\mathcal{R}(t)$ is bounded. (Alternatively, the boundedness can be proved from Gronwall's inequality.)

It remains to show that $\mathcal{R}(t)$ is closed: Let (z_n) be a sequence in $\mathcal{R}(t)$ with $\lim_{n\to\infty} z_n = z$, choose $u_n \in \mathcal{U}$ with $z_n = y(t;u_n)$. By Proposition 2.3 there exists a weakly star convergent subsequence $(u_{n_k})_{k\in\mathbb{N}}$, let $u \in \mathcal{U}$ its limit. We then have

$$\begin{aligned} z &= \lim_{k\to\infty} z_{n_k} = \lim_{k\to\infty} y(t;u_{n_k}) \\ &= \lim_{k\to\infty} \left[\Phi(t,0)y_0 + \int_0^t \Phi(t,s)B(s)u_{n_k}(s)\,ds \right] \\ &= \Phi(t,0)y_0 + \int_0^t \Phi(t,s)B(s)u(s)\,ds \\ &= y(t;u)\,, \end{aligned} \tag{2.17}$$

therefore $z \in \mathcal{R}(t)$. □

Definition 2.5 (Hausdorff distance)
Let $A, B \subset \mathbb{R}^n$, let $x \in \mathbb{R}^n$, sei

$$\delta(x,A) = \inf_{a\in A} \|x-a\|\,. \tag{2.18}$$

The Hausdorff distance of A and B is defined as

$$d(A,B) = \max\{\sup_{x\in B} \delta(x,A)\,,\ \sup_{y\in A} \delta(y,B)\}\,. \tag{2.19}$$

Lemma 2.6 (Hausdorff metric)
Formula (2.19) defines a metric on the set $\mathcal{K}$ of all nonempty compact subsets of $\mathbb{R}^n$, it is called the Hausdorff metric.

Proof: Omitted. □

Proposition 2.7 *The mapping $t \mapsto \mathcal{R}(t)$ is continuous w.r.t. the Hausdorff metric.*

Proof: We first claim that

$$|y(t;u) - y(\tau;u)| \le L(t)|t-\tau| \quad \text{for all } u \in \mathcal{U},\ \tau \le t, \tag{2.20}$$

where the Lipschitz constant $L(t)$ does not depend on u. Indeed, we have

$$y(t;u) - y(\tau;u) = \int_\tau^t A(s)y(s;u) + B(s)u(s)\,ds \tag{2.21}$$

so (2.20) follows from Proposition 2.4 and since $\mathcal{U}$ is bounded.

Let now $\xi \in \mathcal{R}(t)$ and $\tau \le t$ be arbitrary, let $u \in \mathcal{U}$ with $\xi = y(t;u)$. Due to (2.20) we have

$$\delta(\xi, \mathcal{R}(\tau)) \le |y(t;u) - y(\tau;u)| \le L(t)|t-\tau|\,,$$

therefore

$$\sup_{\xi\in\mathcal{R}(t)} \delta(\xi, \mathcal{R}(\tau)) \le L(t)|t-\tau|\,.$$

Interchanging the roles of t and τ yields the assertion. □

Corollary 2.8 *Let $\xi \in \operatorname{int} \mathcal{R}(t)$, $t > 0$. Then there exists an ε-ball V around ξ and an $\eta > 0$ such that*

$$|t - \tau| < \delta \quad \Rightarrow \quad V \subset \operatorname{int}(\mathcal{R}(\tau)). \tag{2.22}$$

Proof: Exercise.

Definition 2.9 *Consider Problem 2.1. A time $T \geq 0$ is called **minimal**, if $y_1 \in \mathcal{R}(T)$ and $y_1 \notin \mathcal{R}(t)$ for all $t < T$. A control $u \in \mathcal{U}$ is called **time optimal**, if $y(T; u) = y_1$.* □

Proposition 2.10 (Existence of time optimal controls)
Let $y_1 \in \mathbb{R}^n$ such that there exists a $t > 0$ with $y_1 \in \mathcal{R}(t)$. Then there exists a minimal $T \geq 0$ such that $y_1 \in \mathcal{R}(T)$.

Proof: Sei $t_n \downarrow T := \inf\{t : t > 0, y_1 \in \mathcal{R}(t)\}$ be a sequence such that $y_1 \in \mathcal{R}(t_n)$. We have

$$0 \leq \delta(y_1, \mathcal{R}(T)) \leq d(\mathcal{R}(t_n), \mathcal{R}(T)) \to 0 \tag{2.23}$$

by Proposition 2.7. Since $\mathcal{R}(T)$ is compact by Proposition 2.4, $y_1 \in \mathcal{R}(T)$. □

Definition 2.11 *A control $u \in \mathcal{U}$ is called **extremal** at time t, if $y(t; u) \in \partial \mathcal{R}(t)$.*

Proposition 2.12 *Every time optimal control is extremal at the minimal time T.*

Proof: Let $u \in \mathcal{U}$ a time optimal control, corresponding to the minimal time T. Then $y(T; u) = y_1 \in \mathcal{R}(T) \subset \operatorname{int}(\mathcal{R}(T)) \cup \partial \mathcal{R}(T)$. We argue by contradiction. If $y(T; u) \in \operatorname{int} \mathcal{R}(T)$, due to Corollary 2.8 there exists $\tau < T$ such that $y_1 \in \operatorname{int}(\mathcal{R}(\tau)) \subset \mathcal{R}(\tau)$. But then T is not the minimal time, a contradiction. Therefore $y(T; u) \in \partial \mathcal{R}(T)$. □

We state the following basic result from convex optimization.

Proposition 2.13 (Separation in $\mathbb{R}^n$)
Let $C \subset \mathbb{R}^n$ be convex, closed and nonempty, let $y \in \mathbb{R}^n$. Then we have

$$y \notin C \quad \Leftrightarrow \quad \exists\, z \in \mathbb{R}^n,\ \|z\|_2 = 1, \text{ with } z^T x < z^T y \text{ for all } x \in C. \tag{2.24}$$

$$y \in \partial C \quad \Rightarrow \quad \exists\, z \in \mathbb{R}^n,\ \|z\|_2 = 1, \text{ with } z^T x \leq z^T y \text{ for all } x \in C. \tag{2.25}$$

□

Above it is not assumed that the interior of C is nonempty.

Proposition 2.14 (Characterization of extremal controls)
*Let $u \in \mathcal{U}$ with associated state $y(\cdot; u)$, let $T > 0$. Then u is extremal at time T if and only if there exists a solution $p : [0, T] \to \mathbb{R}^n$, $p \neq 0$, of the so-called **adjoint system***

$$\dot{p} = -A(t)^T p, \tag{2.26}$$

which satisfies

$$p(T)^T x \leq p(T)^T y(T; u) \quad \text{for all } x \in \mathcal{R}(T), \tag{2.27}$$

as well as

$$p(t)^T B(t) u(t) = \max_{\omega \in \Omega} p(t)^T B(t) \omega \quad \text{for almost all } t \in [0, T]. \tag{2.28}$$

Note that this characterization applies to time optimal controls, since every time optimal control is extremal at time T, by Proposition 2.12.

Proof: If u is not extremal at time T, we have $y(T;u) \in \operatorname{int} \mathcal{R}(T)$. The inequality (2.27) then implies that $p(T) = 0$; due to (2.26) we then must have $p = 0$. Conversely, let u be extremal at time T, that is, $y(T;u) \in \partial\mathcal{R}(T)$. According to Proposition 2.13 we choose a $z \in \mathbb{R}^n$ mit $\|z\|_2 = 1$ und

$$z^T x \le z^T y(T;u) \quad \text{for all } x \in \mathcal{R}(T). \tag{2.29}$$

Let $p : [0,T] \to \mathbb{R}^n$ be the unique solution of (2.26) for the "initial" value (in the sense of an "initial value problem")

$$p(T) = z\,. \tag{2.30}$$

We then have (2.27) by construction. It remains to show that (2.28) holds. For every $t \in [0,T]$ we have

$$\begin{aligned} p(t)^T y(t;u) - p(0)^T y_0 &= \int_0^t \frac{\mathrm{d}}{\mathrm{d}s} \left(p(s)^T y(s;u)\right)\, ds \\ &= \int_0^t \left(\dot p(s)^T y(s;u) + p(s)^T \dot y(s;u)\right)\, ds \\ &= \int_0^t p(s)^T B(s) u(s)\, ds\,. \end{aligned} \tag{2.31}$$

For every $\tilde u \in \mathcal{U}$ we have $y(T;\tilde u) \in \mathcal{R}(T)$. Therefore, (2.27) and (2.31) imply that

$$\begin{aligned} \int_0^T p(s)^T B(s)\tilde u(s)\, ds &= p(T)^T y(T;\tilde u) - p(0)^T y_0 \\ &\le p(T)^T y(T;u) - p(0)^T y_0 \\ &= \int_0^T p(s)^T B(s) u(s)\, ds\,, \quad \text{for all } \tilde u \in \mathcal{U}. \end{aligned} \tag{2.32}$$

We now choose a function $\tilde u : [0,T] \to \Omega$ such that

$$p(t)^T B(t)\tilde u(t) = \max_{\omega\in\Omega} p(t)^T B(t)\omega\,, \quad \text{for a.a. } t \in [0,T]. \tag{2.33}$$

(This is possible for any given t since Ω is compact.) Then we have

$$p(t)^T B(t)(\tilde u(t) - u(t)) \ge 0\,, \quad \text{for a.e. } t \in [0,T]. \tag{2.34}$$

On the other hand, if $\tilde u$ is measurable, then by (2.32)

$$\int_0^T p(t)^T B(t)(\tilde u(t) - u(t))\, dt \le 0\,. \tag{2.35}$$

Then we have equality in (2.34), and the assertion follows. (So far, this argument is not complete, due to the "if $\tilde u$ is measurable". See the following remark.) □

Remark 2.15
(i) Let us consider the special case of a scalar control, that is, $m = 1$. Then Ω is a closed interval, $\Omega = [u_{min}, u_{max}]$. The matrix $B(t)$ has only one column. The function

$$S(t) = p(t)^T B(t)\,, \quad S : [0,T] \to \mathbb{R},$$

is called the **switching function**. The function

$$\tilde{u}(t) = \begin{cases} u_{max}\,, & S(t) > 0\,, \\ u_{min}\,, & S(t) < 0\,, \\ c\,, & S(t) = 0\,, \end{cases} \tag{2.36}$$

where $c \in \Omega$ is arbitrary, satisfies (2.33) and is measurable.
(ii) The problem "find a measurable function f such that $f(t) \in F(t)$ for given sets $F(t)$" is called the problem of **measurable selection**.
(iii) One can circumvent the problem of measurable selection if, following (2.32), one uses a different proof. For example, in points t in which u and B are continuous, one can argue as follows. If there exists an $\omega \in \Omega$ such that

$$p(t)^T B(t)u(t) < p(t)^T B(t)\omega\,,$$

the control

$$\tilde{u}(s) = \begin{cases} \omega\,, & |s-t| < \varepsilon\,, \\ u(s)\,, & |s-t| \geq \varepsilon\,, \end{cases}$$

violates (2.32) for small ε. For a general argument, one has to use the theorem of Lusin from real analysis.

Example 2.16
We consider $\ddot{y} = u$ with an initial condition $y(0) = y_0$ and terminal condition $y(T) = y_1$, and the control constraint $|u(t)| \leq 1$. We have

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\,, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}\,, \quad \Omega = [-1,1]\,. \tag{2.37}$$

The adjoint system is given by

$$\dot{p}_1 = 0\,, \tag{2.38}$$

$$\dot{p}_2 = -p_1\,. \tag{2.39}$$

The maximum condition becomes

$$p(t)^T B u(t) = \max_{\omega \in [-1,1]} p(t)^T B\omega \qquad \text{a.e. in } t. \tag{2.40}$$

The switching function is given by

$$S(t) = p(t)^T B = p_2(t)\,, \tag{2.41}$$

and the time optimal control is

$$u(t) = \begin{cases} 1\,, & p_2(t) > 0\,, \\ -1\,, & p_2(t) < 0\,. \end{cases} \tag{2.42}$$

The adjoint system can be solved explicitly,

$$p_1(t) = \alpha\,, \quad p_2(t) = \beta - \alpha t\,, \tag{2.43}$$

with some constants α, β which cannot both be zero due to Proposition 2.14. The switching function $S = p_2$ therefore has at most one zero t_*. There are two possibilites for the structure of the time optimal control u: It can be constant (either 1 or -1), or it has exactly one switching point t_* where it switches from -1 to 1 or vice versa. Which of these cases occurs, depends on the values of y_0 and y_1.

For the example from the introduction,

$$y_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\,, \quad y_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}\,, \tag{2.44}$$

the only possibility is

$$u(t) = \begin{cases} 1\,, & t < t_*\,, \\ -1\,, & t > t_*\,. \end{cases} \tag{2.45}$$

The values $T = 2$ and $t_* = 1$ can then be computed from the state equation and (2.44).

In the previous example, the time optimal control is piecewise constant, and takes as values only the end points ± 1 of the admissible set $\Omega = [-1, 1]$. The task of finding the optimal control is reduced to finding out its "structure" (that is, the successive values taken by the control) and the exact location of the switching points. We now look more closely into situations where this might occur.

Constant coefficients, polyhedral control constraint. From now on we assume that

A and B are constant (no dependence on time)

Ω is a convex polyhedron (that is, defined by finitely many linear inequalities)

Let $u : [0, T] \to \mathbb{R}^m$ be a time optimal control, let $p : [0, T] \to \mathbb{R}^n$ be the adjoint obtained from Proposition 2.14. Let $S : [0, T] \to \mathbb{R}^m$ denote the **switching function**

$$S(t) = B^T p(t)\,. \tag{2.46}$$

S is continuous since p is continuous. The maximum condition then becomes

$$S(t)^T u(t) = \max_{\omega \in \Omega} S(t)^T \omega\,. \tag{2.47}$$

For each fixed t, $u(t)$ thus maximizes the linear functional $\omega \mapsto S(t)\omega$ over the polyhedron Ω. There are two possibilities:

- (Regular case:) The maximum is unique, $u(t)$ is a vertex of Ω.
- (Singular case:) The maximum is attained at a face F of Ω, which can be 1 D (an edge) or higher dimensional. The vector $S(t)$ is orthogonal to this face, and

  $$S(t)^T w = 0 \tag{2.48}$$

  holds for all edges w of F,

  $$w = v_1 - v_2\,, \tag{2.49}$$

  if the edge connects the vertices v_1 and v_2.

Let t be a time where the maximum is unique. Since S is continuous, the maximizing vertex does not change near t, so $u(\tau) = u(t)$ if $|\tau - t|$ is sufficiently small. Let I_t be the maximal interval such that $u(\tau) = u(t)$ for all $\tau \in I_t$. If a boundary point of I_t belongs to the interior $(0, T)$ of the whole time interval, the maximum is not unique at t, and $S(t)^T w = 0$ for some edge w of Ω.

Structure of the switching function. Let us look at the function

$$s_w(t) = S(t)^T w = p(t)^T B w\,, \quad w \text{ edge of } \Omega. \tag{2.50}$$

The function p solves $\dot{p} = -A^T p$, a linear differential equation with constant coefficients. In this case, we have the explicit solution formula

$$p(t) = e^{(T-t)A^T} p(T)\,, \tag{2.51}$$

and thus

$$s_w(t) = p(t)^T B w = p(T)^T e^{(T-t)A} B w\,. \tag{2.52}$$

This function is defined not only on $[0, T]$, but can be extended to the complex plane,

$$\tilde{s}_w(z) = p(T)^T e^{(T-z)A} B w\,. \tag{2.53}$$

The function $\tilde{s}_w : \mathbb{C} \to \mathbb{C}$ is holomorphic (= differentiable in the sense of complex analysis) on $\mathbb{C}$. By the identity theorem of complex analysis, for the behaviour of $\tilde{s}_w$ (and thus, of s_w) on $[0, T]$ there are only two possibilities:

$$\tilde{s}_w \text{ has finitely many zeroes on } [0, T], \tag{2.54}$$

or

$$\tilde{s}_w = 0 \text{ in } [0, T]. \tag{2.55}$$

Definition 2.17
*Let Ω be a polyhedron. A control $u : [0, T] \to \mathbb{R}^m$ is called a **bang-bang control** for Problem 2.1, if u is piecewise constant and $u(t)$ is a vertex of Ω for all $t \in (0, T)$.* □

Proposition 2.18 *Assume that A, B are constant and Ω is a polyhedron. Let $u : [0, T] \to \mathbb{R}^m$ be a time optimal control such that (2.54) holds for all edges w of Ω. Then u is a bang-bang control.*

Proof: We define the switching set

$$P = \{t : t \in [0, T]\,,\ s_w(t) = 0 \text{ for some edge } w \text{ of } \Omega\}\,. \tag{2.56}$$

Since Ω has only finitely many edges, P is a finite set by assumption (2.54). On $[0, T] \setminus P$, all functions s_w are nonzero, therefore we are in the regular case. Thus, on $[0, T] \setminus P$ the time optimal control u is locally constant and has vertices of Ω as its values. Since P is finite, u has at most finitely many discontinuities. □

Let us consider the case (2.55), $s_w = 0$ on $[0,T]$ for some edge w of Ω. Then all derivatives of s_w are zero, and since $\dot{p}(t)^T = -p(t)^T A$, we have

$$0 = s_w^{(k)}(t) = (-1)^k p(t)^T A^k B w\,, \quad \text{for all } k \in \mathbb{N},\, t \in (0,T). \tag{2.57}$$

Since $p \neq 0$, the n vectors

$$Bw\,,\; ABw\,,\; \ldots\,,\; A^{n-1}Bw$$

must be linearly dependent.

Definition 2.19 *We say that (A,B,Ω) satisfies the **normality condition**, if the vectors*

$$Bw, ABw, \ldots, A^{n-1}Bw \tag{2.58}$$

are linearly independent for every edge w of Ω. □

Proposition 2.20 *Let (A,B,Ω) satisfy the normality condition. Then every time optimal control is bang-bang.*

Proof: The arguments above Definition 2.19 show that if the normality condition holds, then s_w has only finitely many zeroes for every edge w of Ω. The claim now follows from Proposition 2.18. □

In the scalar case $m = 1$, $\Omega = [u_{min}, u_{max}]$ there is exactly one edge w, namely $u_{max} - u_{min}$, and $A^k B \in \mathbb{R}^n$ for all $k \geq 0$. The normality condition then becomes

$$\text{The matrix with columns } B\; AB\; \ldots A^{n-1}B \text{ is invertible.} \tag{2.59}$$

In control theory it is shown that this is the case if and only if the system $\dot{y} = Ay + Bu$ is **controllable**, that is, for every $y_0, y_1 \in \mathbb{R}^n$ and $T > 0$ there exists a control u such that $y(0) = y_0$ and $y(T) = y_1$.

3 Existence of Optimal Controls

Let X be a vector space, $K \subset X$ convex and nonempty, $J : X \to \mathbb{R}$ convex. We consider the problem of convex optimization

$$\text{minimize } J(u)\,, \quad u \in K\,. \tag{3.1}$$

We say that $u \in X$ is a **solution** (or a **minimizer**) of (3.1), if $u \in K$ and $J(u) \le J(w)$ for all $w \in K$. A sequence $\{u_n\}$ in K is called a **minimizing sequence**, if $J(u_n) \to \inf\{J(u) : u \in K\}$.

In order that (3.1) makes sense, it is not required that J is defined on all of X; it suffices that $J : K \to \mathbb{R}$.

In the theory for (3.1), the requirement that J is defined on all of X does not impose a restriction; if $J : K \to \mathbb{R}$, we may extend J to all of X by setting

$$\tilde{J}(u) = \begin{cases} J(u)\,, & u \in K\,, \\ +\infty\,, & u \notin K\,. \end{cases} \tag{3.2}$$

Replacing J by $\tilde{J}$ in (3.1) does not change the set of solutions.

When we solve (3.1) numerically, however, if the algorithm uses arguments v which are not elements of K, it plays a role how J is defined outside of K.

When $X = \mathbb{R}^n$, and K is defined by finitely many constraints (inequalities, equations), we are in the realm of finite-dimensional optimization. When either $\dim X = \infty$ or K is defined by infinitely many constraints, one speaks of semi-infinite optimization. If both $\dim X = \infty$ and there are an infinite number of constraints, the optimization problem is called infinite-dimensional. Optimal control problems are infinite-dimensional. A typical function space for the controls is the Hilbert space

$$X = L^2(\Omega)\,, \quad \Omega \subset \mathbb{R}^n\,,$$

a space which also includes discontinuous control functions. Another typical control space is $L^\infty(\Omega)$.

Definition 3.1
Let X be a normed space, $K \subset X$, $J : K \to \mathbb{R}$.
*J is called **lower semicontinuous on K**, if*

$$J(u) \le \liminf_{n\to\infty} J(u_n) \tag{3.3}$$

holds for every sequence $\{u_n\}$ in K which converges to some $u \in K$.
*J is called **coercive on K**, if $J(u_n) \to \infty$ for every sequence $\{u_n\}$ in K with $\|u_n\| \to \infty$.*

The definition of "coercive" is not uniform in the literature.

If K is bounded, then J is coercive on K because there is no unbounded sequence in K.

The following proposition contains the typical ingredients of an existence result in infinite-dimensional optimization. We first formulate it in finite dimensions.

Proposition 3.2 *Let $X = \mathbb{R}^n$, $K \subset X$ closed and nonempty, $J : K \to \mathbb{R}$ lower semicontinuous, bounded from below and coercive on K. Then there exists a minimizer for J on K. If moreover J is strictly convex, the minimizer is unique.*

Proof: Since K is nonempty and J is bounded from below,

$$m = \inf_{u \in K} J(u)$$

is a real number (that is, neither $-\infty$ nor $+\infty$). Let $\{u_n\}$ be a minimizing sequence in K. Since J is coercive, $\{u_n\}$ is bounded (otherwise there would exist a subsequence $\{u_{n_k}\}$ in K with $\|u_{n_k}\| \to \infty$ and thus $J(u_{n_k}) \to \infty$, a contradiction). Since X is finite-dimensional, there exists a subsequence $\{u_{n_k}\}$ with $u_{n_k} \to u$ for some $u \in X$. Since K is closed, $u \in K$. Since J is lower semicontinuous,

$$J(u) \le \liminf_{k \to \infty} J(u_{n_k}) = m = \inf_{u \in K} J(u)\,.$$

Therefore, u is a minimizer of J on K. If J is strictly convex,

$$J\Big(\frac{u_1 + u_2}{2}\Big) < \frac{1}{2} J(u_1) + \frac{1}{2} J(u_2)$$

holds if $u_1 \neq u_2$. Thus, there cannot exist two different minimizers. □

If we restrict our attention to $X = \mathbb{R}^n$, the proof can be formulated more concisely. Coercivity of J then implies that the sublevel sets of admissible points,

$$\{w : J(w) \le \alpha\} \cap K$$

are nonempty (if $\alpha > \inf_K J$), closed and bounded, thus compact. Therefore, on such a set J attains a minimum u, which is a minimizer of the original problem.

The arguments above work because every minimizing sequence turns out to have a convergent subsequence, whose limit then is a minimizer. When X is infinite-dimensional, closed bounded subsets of X need not be compact, and a minimizing sequence does not necessarily have a subsequence which converges in the norm of X. Instead of compactness, one then employs weak compactness. This yields weakly convergent minimizing sequences. In order that "everything else works as before" we additionally assume that J and K are convex.

Let X be a Hilbert space with a scalar product $\langle \cdot, \cdot \rangle$. A sequence $\{u_n\}$ in X **converges weakly** to $u \in X$, denoted by $u_n \rightharpoonup u$, if

$$\lim_{n \to \infty} \langle u_n, v \rangle = \langle u, v \rangle \ , \quad \text{for all } v \in X. \tag{3.4}$$

If $X = \mathbb{R}^n$, convergence and weak convergence coincide.

We need the following two results from functional analysis.

Proposition 3.3 *Let X be a Hilbert space, $K \subset X$ convex and closed, $\{u_n\}$ a sequence in K, $u_n \rightharpoonup u$ for some $u \in X$. Then $u \in K$.*

Proof: This is a result from functional analysis. □

Proposition 3.4 *Let X be a Hilbert space, $\{u_n\}$ be a bounded sequence in X. Then $\{u_n\}$ has a weakly convergent subsequence.*

Proof: This is a result from functional analysis. □

A functional $J : K \to \mathbb{R}$, $K \subset X$, is called **weakly lower semicontinuous**, if

$$J(u) \leq \liminf_{n\to\infty} J(u_n) \tag{3.5}$$

for every sequence $\{u_n\}$ in K with $u_n \rightharpoonup u$.

Proposition 3.5 *Let X be a Hilbert space, $K \subset X$ convex and closed, $J : K \to \mathbb{R}$ convex and lower semicontinuous on K. Then J is weakly lower semicontinuous on K.*

Proof: For arbitrary $\alpha \in \mathbb{R}$, consider the set

$$K_\alpha = \{w : J(w) \leq \alpha\} \cap K\,.$$

It is convex since K and J are convex; it is closed since for every sequence $\{u_n\}$ with $u_n \in K$, $J(u_n) \leq \alpha$ and $u_n \to u$ we have $u \in K$ and

$$J(u) \leq \liminf_{n\to\infty} J(u_n) \leq \alpha\,.$$

Now assume that J is not weakly lower semicontinuous on K. Then there exists a sequence $\{u_n\}$ in K with $u_n \rightharpoonup u$ and

$$J(u) > \liminf_{n\to\infty} J(u_n)\,.$$

Choose α such that

$$J(u) > \alpha > \liminf_{n\to\infty} J(u_n)\,. \tag{3.6}$$

Let $\{u_{n_k}\}$ be a subsequence such that $J(u_{n_k}) \leq \alpha$ for all k. Since $u_{n_k} \rightharpoonup u$ and K_α is closed and convex, we have that $u \in K_\alpha$ by Proposition 3.3. Thus $J(u) \leq \alpha$, a contradiction to (3.6). Therefore, J is weakly semicontinuous on K. □

Proposition 3.6 *Let X be a Hilbert space, $\emptyset \neq K \subset X$ closed and convex, $J : K \to \mathbb{R}$ convex, lower semicontinuous, bounded from below and coercive on K. Then there exists a minimizer for J on K. If moreover J is strictly convex, the minimizer is unique.*

Proof: The proof parallels that of Proposition 3.2, taking into account the properties of weak convergence. As before, one checks that

$$m = \inf_{u\in K} J(u)$$

is a real number, and that every minimizing sequence $\{u_n\}$ is bounded. By Proposition 3.4 there exists a subsequence $\{u_{n_k}\}$ with $u_{n_k} \rightharpoonup u$ for some $u \in X$. Since K is closed and convex, $u \in K$ by Proposition 3.3. Moreover, J is weakly lower semicontinuous by Proposition 3.5, thus

$$J(u) \leq \liminf_{k\to\infty} J(u_{n_k}) = m = \inf_{u\in K} J(u)\,,$$

and u is a minimizer of J on K. Strict convexity of J implies uniqueness as in Proposition 3.2. □

Application to ordinary differential equations. Let us consider the linear ODE system

$$\dot{y} = Ay + Bu\,, \quad y(0) = y_0\,, \tag{3.7}$$

for $y : [0,T] \to \mathbb{R}^n$, $u : [0,T] \to \mathbb{R}^m$, $A \in \mathbb{R}^{(n,n)}$, $B \in \mathbb{R}^{(n,m)}$, $y_0 \in \mathbb{R}^n$. From the theory of ODE's we have the solution formula

$$y(t) = e^{tA}y_0 + e^{tA}\int_0^t e^{-sA}Bu(s)\,ds\,, \quad t \in [0,T]\,. \tag{3.8}$$

We want to take controls u in the space $X = L^2(0,T)$. The integrand is an element of $L^2(0,T)$ because

$$\sup_{s\in[0,T]} \|e^{-sA}B\| < \infty\,. \tag{3.9}$$

By a result of integration theory, the function

$$t \mapsto \int_0^t f(s)ds\,,$$

is continuous (even absolutely continuous) if f is integrable. Thus, (3.8) defines a function $y \in C[0,T]$. The corresponding operator

$$S : L^2(0,T) \to C[0,T]\,, \quad y = Su\,, \tag{3.10}$$

is called the **control-to-state mapping**. It is affine linear, that is,

$$\tilde{S}u = Su - S0 \tag{3.11}$$

defines a linear operator $\tilde{S} : L^2(0,T) \to C[0,T]$. Moreover, $\tilde{S}$ (and hence S) is continuous, since

$$\begin{aligned}\|\tilde{S}u\|_\infty &= \max_{t\in[0,T]} \|(Su)(t) - (S0)(t)\| \le C_T \int_0^T \|u(s)\|\,ds \\ &\le \sqrt{T}\sqrt{\int_0^T \|u(s)\|^2\,ds} = C_T\sqrt{T}\|u\|_{L^2(0,T)}\end{aligned}$$

for some constant C_T which does not depend on u, obtained via (3.8) and (3.9).

We also consider a control constraint

$$u(t) \in U_{ad} \quad \text{a.e. in } (0,T), \quad U_{ad} \subset \mathbb{R}^m \text{ closed, convex and nonempty,} \tag{3.12}$$

and define the corresponding set K by

$$K = \{u : u \in L^2(0,T)\,,\, u(t) \in U_{ad} \text{ for a.a. } t \in (0,T).\} \tag{3.13}$$

The set K is a closed, convex and nonempty subset of $L^2(0,T)$.

We now consider the problem to

$$\text{minimize } J(u) = J_1(Su) + J_2(u), \quad \text{subject to} \quad u \in K\,. \tag{3.14}$$

Here,

$$J_1 : C[0,T] \to \mathbb{R}\,, \quad J_2 : L^2(0,T) \to \mathbb{R}\,. \tag{3.15}$$

An example is given by

$$J_1(y) = \frac{1}{2}\int_0^T \|y(t) - y_d(t)\|^2\,dt\,, \quad y_d \in L^2(0,T) \text{ given}, \tag{3.16}$$

$$J_2(u) = \frac{\alpha}{2}\int_0^T \|u(t)\|^2\,dt\,, \quad \alpha \geq 0 \text{ given}. \tag{3.17}$$

The functional J_2 is coercive on K if $\alpha > 0$, or if U_{ad} is bounded. In this example, J_1 and J_2 are quadratic, while the differential equation is linear (and, hence, the control-to-state mapping S is affine linear). Such control problems are called **linear-quadratic**.

Proposition 3.7 *Consider the optimal control problem (3.14), where S is the control-to-state mapping from (3.10) and K is the control constraint (3.13). Assume that J_1 and J_2 are convex, continuous and bounded from below, and that J_2 is coercive on K. Then the problem has a solution $u \in K$.*

Proof: We check that the assumptions of Proposition 3.6 are satisfied. The cost functional J is convex since S is affine linear and J_1 and J_2 are convex. J is bounded from below since so are J_1 and J_2. J is coercive since J_2 is coercive and J_1 is bounded from below. J is continuous since J_1, J_2 and S are continuous. We already have stated above that K is closed, convex and nonempty. □

If one adds a terminal constraint to the problem,

$$y(T) = y_1\,,$$

one may subsume this under Problem (3.14), modifying K to

$$K = \{u : u \in L^2(0,T)\,,\ u(t) \in U_{ad} \text{ for a.a. } t \in (0,T),\ (Su)(T) = y_1.\}$$

This modified set K is still convex and closed, but in order to check whether it is nonempty, one has to consider the controllability problem.

Application to an elliptic problem. We consider the following optimal control problem. Let $\Omega \subset \mathbb{R}^d$ be open and bounded. The state $y : \Omega \to \mathbb{R}$ should satisfy

$$\begin{aligned} -\Delta y &= \beta u \quad && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{3.18}$$

Here, $u : \Omega \to \mathbb{R}$ is the control function, and $\beta \in L^\infty(\Omega)$ is given. We want to minimize

$$\frac{1}{2}\int_\Omega (y(x) - y_d(x))^2\,dx + \frac{c}{2}\int_\Omega u(x)^2\,dx\,. \tag{3.19}$$

In contrast to the problem in Chapter 1, the control acts in the interior of Ω instead of on the boundary; this is mathematically simpler to treat.

The question, whether and why the boundary value problem (3.18) has a unique solution y for given u, is discussed in the theory of partial differential equations. Here, we present a brief summary. In the variational approach, which we use here, (3.19) is replaced by a variational equation. Let $\phi \in C_0^\infty(\Omega)$ be a so-called test function. Multiplication and partial integration gives

$$\int_\Omega \beta u \varphi \, dx = \int_\Omega (-\Delta y)\varphi \, dx = \int_\Omega \langle \nabla y, \nabla \varphi \rangle \, dx \,.$$

(There are no boundary terms since φ is zero on the boundary.) One says that $y : \Omega \to \mathbb{R}$ is a **weak solution** of (3.18), if

$$\int_\Omega \langle \nabla y, \nabla \varphi \rangle \, dx = \int_\Omega \beta u \varphi \, dx \,, \quad \text{for all } \varphi \in C_0^\infty(\Omega). \tag{3.20}$$

When $u \in L^2(\Omega)$ and β is bounded, this makes sense if $y \in L^2(\Omega)$ as well as $\nabla y \in L^2(\Omega)$ (that is, all partial derivatives $\partial_i y$ are in $L^2(\Omega)$). Moreover, the function space for y should be a Hilbert space. For this reason, for the partial derivatives $\partial_i y$ one uses the concept of a **weak derivative**. One again starts from a partial integration formula, namely

$$\int_\Omega y \, \partial_i \varphi \, dx = -\int_\Omega \partial_i y \, \varphi \, dx \,.$$

A function $z \in L^2(\Omega)$ is called the weak i-th partial derivative of $y \in L^2(\Omega)$, and denoted by $\partial_i y$, if

$$\int_\Omega y \, \partial_i \varphi \, dx = -\int_\Omega z \, \varphi \, dx \,. \tag{3.21}$$

One then defines the function space

$$H^1(\Omega) = \{ y : y \in L^2(\Omega) \,,\, \partial_i y \in L^2(\Omega) \text{ for all } i \} \,. \tag{3.22}$$

This function space is a Hilbert space with the scalar product and norm

$$\begin{aligned} \langle y, v \rangle &= \sum_{i=1}^n \int_\Omega \partial_i y \, \partial_i v \, dx + \int_\Omega y \, v \, dx \,, \\ \|y\|^2_{H^1(\Omega)} &= \sum_{i=1}^n \int_\Omega (\partial_i y)^2 \, dx + \int_\Omega y^2 \, dx \,. \end{aligned} \tag{3.23}$$

To incorporate the boundary condition $y = 0$, one uses the subspace

$$H_0^1(\Omega) = \overline{C_0^\infty(\Omega)} \,,$$

the closure of the space of test functions w.r.t. the norm of $H^1(\Omega)$. This gives rise to the norm

$$\|y\|^2_{H_0^1(\Omega)} = \int_\Omega \langle \nabla y, \nabla y \rangle \, dx \,, \tag{3.24}$$

which on $H_0^1(\Omega)$ is equivalent to the norm defined by (3.23).

Proposition 3.8 *Let $\beta \in L^\infty(\Omega)$. For every $u \in L^2(\Omega)$ there exists a unique solution $y \in H_0^1(\Omega)$ of (3.20). Moreover, there exists $C > 0$ such that*

$$\|y\|_{H_0^1(\Omega)} \leq C\|u\|_{L^2(\Omega)}\,. \tag{3.25}$$

Proof: This is a result from the theory of partial differential equations. □

Thus, the control-to-state mapping $u \mapsto y =: Su$ is well-defined.

Corollary 3.9 *The control-to-state mapping S for (3.18) is a linear and continous mapping from $L^2(\Omega)$ to $H_0^1(\Omega)$.* □

The optimal control problem now has the form: Minimize

$$J(u) = J_1(Su) + J_2(u) = \frac{1}{2}\int_\Omega ((Su)(x) - y_d(x))^2\,dx + \frac{c}{2}\int_\Omega u(x)^2\,dx\,, \quad c \geq 0\,, \tag{3.26}$$

subject to

$$u \in K = \{u : u \in L^2(\Omega)\,, \quad u(t) \in U_{ad} \text{ for a.a. } t \in (0,T)\}\,. \tag{3.27}$$

We consider two cases:

$$\text{(a)} \quad c > 0\,,\ U_{ad} = \mathbb{R}\,, \qquad \text{(b)} \quad c \geq 0\,,\ U_{ad} = [u_{min}, u_{max}]\,. \tag{3.28}$$

As in the subsection on the ODE control problem, this is a linear-quadratic problem. The cost functional J is convex, coercive in both cases of (3.28), and bounded from below by 0. It is also continuous on $L^2(\Omega)$, since J_2 is obviously continuous on $L^2(\Omega)$ and J_1 is a composition of continuous mappings

$$L^2(\Omega) \to H_0^1(\Omega) \to L^2(\Omega) \to \mathbb{R}\,, \quad u \mapsto Su \mapsto Su \mapsto J_1(Su)\,.$$

Moreover, K is a closed convex subset of $L^2(\Omega)$ in both cases of (3.28).

Proposition 3.10 *The optimal control problem (3.26) has a solution $u \in L^2(\Omega)$. In the case $c > 0$, the solution is unique.*

Proof: This is a consequence of Proposition 3.6. In the case $c > 0$, J is strictly convex. □

4 Adjoint Systems and Optimality

Let X be a vector space, $K \subset X$, $j : X \to \mathbb{R}$. We consider

$$\text{minimize } j(u)\,, \quad u \in K\,. \tag{4.1}$$

We say that j is **directionally differentiable** at $u \in X$ in the direction $h \in X$, if

$$j'(u;h) = \lim_{\lambda \downarrow 0} \frac{j(u+\lambda h) - j(u)}{\lambda} \tag{4.2}$$

exists. We then call $j'(u;h)$ the directional derivative of j at u in the direction h. If it exists for all $h \in X$, we say that j is directionally differentiable at u.

If $u \in K$ and K is convex, we call h an **admissible direction** for K at u, if there exists $\lambda > 0$ such that $u + \lambda h \in K$. We then have $u + sh \in K$ for every $s \in [0, \lambda]$. We denote

$$K(u) = \{h : h \text{ is an admissible direction for } K \text{ at } u\}\,. \tag{4.3}$$

Proposition 4.1 *Let X be a vector space, $K \subset X$ convex, let u be a minimizer for (4.1) and j be directionally differentiable at u. Then*

$$j'(u;h) \geq 0\,, \quad \textit{for all } h \in K(u). \tag{4.4}$$

Proof: For every $h \in K(u)$, we have $0 \leq (j(u+\lambda h) - j(u))/\lambda$, if $\lambda > 0$ is sufficiently small. Passing to the limit $\lambda \to 0$ we obtain the assertion. □

Example 4.2
Let $X = L^2(\Omega)$, $f \in L^2(\Omega)$,

$$j(u) = \frac{1}{2}\|u - f\|_2^2 = \frac{1}{2}\int_\Omega (u(x) - f(x))^2\,dx\,, \quad u \in L^2(\Omega)\,. \tag{4.5}$$

Then for $h \in L^2(\Omega)$

$$\begin{aligned} j(u+\lambda h) - j(u) &= \frac{1}{2}\int_\Omega (u + \lambda h - f)^2 - (u-f)^2\,dx \\ &= \frac{1}{2}\int_\Omega (u-f)^2 + 2\lambda(u-f)h + \lambda^2 h^2 - (u-f)^2\,dx \end{aligned}$$

and

$$j'(u;h) = \int_\Omega (u(x) - f(x))h(x)\,dx\,. \tag{4.6}$$

If u is a minimizer of j on some convex subset $K \neq \emptyset$ of $L^2(\Omega)$, then u is the projection of f on K. Proposition 4.1 says that

$$0 \leq j'(u;h) = \int_\Omega (u(x) - f(x))h(x)\,dx = \langle u - f, h\rangle\,, \quad \text{for all } h \in K(u),$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $L^2(\Omega)$. This is a particular case of the projection theorem in Hilbert space. □

In the example above, the mapping $h \mapsto j'(u;h)$ is a functional on X, that is, a linear continuous mapping from $X = L^2(\Omega)$ to $\mathbb{R}$, it has the form

$$h \mapsto \langle u - f, h \rangle \ .$$

This functional is called the **gradient** of j at u,

$$(\nabla j(u))(h) = \langle u - f, h \rangle \ . \tag{4.7}$$

Since $X = L^2(\Omega)$ is a Hilbert space, we can "identify" it with the element $u - f$ of X (a special case of the Riesz isomorphism). We then also write

$$\nabla j(u) = u - f \, . \tag{4.8}$$

Problem 4.3 We consider the optimization problem

$$\text{minimize } J(y,u) = J_1(y) + J_2(u) \, , \quad y = Su \, , \quad u \in K \, . \tag{4.9}$$

Here, u stands for the control and y for the state. We assume that X and Y are Hilbert spaces, $J_1 : Y \to \mathbb{R}$, $J_2 : X \to \mathbb{R}$, $K \subset X$ is a closed convex set and $S : X \to Y$ is continuous and affine linear, that is,

$$Su = \tilde{S}u + S0 \, , \quad \tilde{S} : X \to Y \text{ linear and continuous.} \tag{4.10}$$

Setting

$$j(u) = J(Su, u) \, , \tag{4.11}$$

we can reduce (4.9) to (4.1). □

In order to compute the gradient $\nabla j(u)$, we need the chain rule.

A mapping $F : X \to Y$ between normed spaces X and Y is called directionally differentiable at u if the limit

$$F'(u;h) = \lim_{\lambda \downarrow 0} \frac{F(u + \lambda h) - F(u)}{\lambda} \tag{4.12}$$

exists for every $h \in X$. We call $F'(u;h)$ the **directional derivative** of F at u in the direction h. Thus, $F'(u;h) \in Y$.

It follows immediately from the definition that

$$F'(u;th) = tF'(u;h) \, , \quad \text{for all } t \geq 0. \tag{4.13}$$

Lemma 4.4 *Let X, Y be normed spaces, $F : X \to Y$ directionally differentiable and locally Lipschitz continuous. Then the mapping $h \mapsto F'(u;h)$ is locally Lipschitz continuous for every $h \in X$.*

Proof: Let $u \in X$, let L be a Lipschitz constant for F near u. Let $h, \tilde{h} \in X$ be arbitrary. Then

$$\begin{aligned}\Big\| \frac{F(u+\lambda h) - F(u)}{\lambda} - \frac{F(u+\lambda \tilde{h}) - F(u)}{\lambda} \Big\| &= \Big\| \frac{F(u+\lambda h) - F(u+\lambda \tilde{h})}{\lambda} \Big\| \\ &\leq \frac{L\|\lambda h - \lambda \tilde{h}\|}{\lambda} = L\|h - \tilde{h}\|\,,\end{aligned}$$

if $\lambda > 0$ is sufficiently small. Letting $\lambda \to 0$ yields

$$\|F'(u;h) - F'(u;\tilde{h})\| \leq L\|h - \tilde{h}\|\,.$$

□

Proposition 4.5 *Let X, Y, Z be normed spaces, let $F : X \to Y$ and $G : Y \to Z$ be locally Lipschitz continuous and directionally differentiable at u (at $F(u)$, respectively). Then so is $G \circ F$, and*

$$(G \circ F)'(u;h) = G'(F(u); F'(u;h)) \tag{4.14}$$

for all $h \in X$.

Proof: The composition of locally Lipschitz continuous functions is locally Lipschitz continuous, as an immediate consequence of the definition. Furthermore, we have

$$\begin{aligned}&\Big\| \frac{G(F(u+\lambda h)) - G(F(u))}{\lambda} - G'(F(u); F'(u;h)) \Big\| \\ &\quad \leq \quad \Big\| \frac{G(F(u+\lambda h)) - G(F(u) + \lambda F'(u;h))}{\lambda} \Big\| \\ &\qquad + \Big\| \frac{G(F(u) + \lambda F'(u;h))) - G(F(u))}{\lambda} - G'(F(u); F'(u;h)) \Big\|\end{aligned} \tag{4.15}$$

If L_G is a local Lipschitz constant for G, then for $\lambda \to 0$

$$\Big\| \frac{G(F(u+\lambda h)) - G(F(u) + \lambda F'(u;h))}{\lambda} \Big\| \leq L_G \Big\| \frac{F(u+\lambda h) - F(u) - \lambda F'(u;h))}{\lambda} \Big\| \;\to\; 0\,.$$

The third term in (4.15) goes to zero as $\lambda \to 0$ since G is directionally differentiable. □

If $h \mapsto F'(u;h)$ defines a linear continuous mapping from X to Y, we denote it by $DF(u)$. In that case

$$F'(u;h) = DF(u)h\,.$$

If the same is true for G, the chain rule can be written as

$$D(G \circ f)(u) = DG(F(u)) \circ DF(u)\,. \tag{4.16}$$

The more general case where $h \mapsto F'(u;h)$ is not linear typically arises in **nonsmooth optimization**, for example when the cost functional includes terms like

$$\int_\Omega |u(x)|\,dx\,, \quad \int_\Omega |\nabla u(x)|\,dx\,, \quad \sup_{x\in\Omega} |u(x)|\,.$$

We return to the task of computing the derivative of $j(u) = J_1(Su) + J_2(u)$ in the setting of Problem 4.3. We have

$$S'(u;h) = \tilde{S}h\,, \tag{4.17}$$

since S is affine linear, thus the derivative does not depend on u. If J_1 and J_2 are Lipschitz continuous and directionally differentiable, we get from the chain rule

$$j'(u;h) = J_1'(Su;\tilde{S}h) + J_2'(u;h)\,. \tag{4.18}$$

If moreover the derivatives of J_1 and J_2 are linear continuous mappings from Y to $\mathbb{R}$ and from X to $\mathbb{R}$, respectively, then in the Hilbert space setting of Problem 4.3 we get

$$\begin{aligned} \langle \nabla j(u), h\rangle &= \Big\langle \nabla J_1(Su), \tilde{S}h\Big\rangle + \langle \nabla J_2(u), h\rangle \\ &= \Big\langle \nabla_y J(Su,u), \tilde{S}h\Big\rangle + \langle \nabla_u J(Su,u), h\rangle\ . \end{aligned} \tag{4.19}$$

The second line is just an alternative notation, since the derivatives of J_1 and J_2 may also be written as partial derivatives of J.

Application to ordinary differential equations. We now look at the computation of the composite mapping

$$h \mapsto \tilde{S}h \mapsto \Big\langle \nabla J_1(Su), \tilde{S}h\Big\rangle$$

in the case where S represents the solution of a linear system of ordinary differential equations. As before, we consider

$$\dot{y} = Ay + Bu\,, \quad y(0) = y_0\,, \tag{4.20}$$

for $y : [0,T] \to \mathbb{R}^n$, $u : [0,T] \to \mathbb{R}^m$, $A \in \mathbb{R}^{(n,n)}$, $B \in \mathbb{R}^{(n,m)}$, $y_0 \in \mathbb{R}^n$. We recall the solution formula

$$\begin{aligned} y(t) &= e^{tA}y_0 + \int_0^t e^{(t-s)A}Bu(s)\,ds \\ &=: y_{in}(t) + (\tilde{S}u)(t)\,, \quad t \in [0,T]\,. \end{aligned} \tag{4.21}$$

We have seen that $u \mapsto y$ defines an affine-linear continuous control-to-state mapping $S : L^2(0,T) \to C[0,T]$ with linear part $\tilde{S}$. Since the embedding of $C[0,T]$ into $L^2(0,T)$ is continuous,

$$S, \tilde{S} : L^2(0,T) \to L^2(0,T) \tag{4.22}$$

are affine linear (linear, resp.) and continuous.

Setting $v = \nabla J_1(Su)$, we get

$$\begin{aligned}
\Big\langle \nabla J_1(Su), \tilde{S}h \Big\rangle = \Big\langle v, \tilde{S}h \Big\rangle &= \int_0^T \Big\langle v(t), \int_0^t e^{(t-s)A} Bh(s)\, ds \Big\rangle\, dt \\
&= \int_0^T \int_s^T \big\langle v(t), e^{(t-s)A} Bh(s) \big\rangle\, dt\, ds \\
&= \int_0^T \int_s^T \Big\langle e^{(t-s)A^T} v(t), Bh(s) \Big\rangle\, dt\, ds \\
&= \int_0^T \Big\langle \int_s^T e^{(t-s)A^T} v(t)\, dt, Bh(s) \Big\rangle\, ds \\
&= \int_0^T \Big\langle \int_t^T e^{(s-t)A^T} v(s)\, ds, Bh(t) \Big\rangle\, dt \\
&= \int_0^T \Big\langle \int_t^T e^{(s-t)A^T} \nabla J_1(Su)(s)\, ds, Bh(t) \Big\rangle\, dt\,.
\end{aligned} \tag{4.23}$$

Setting

$$p(t) = \int_t^T e^{(s-t)A^T} \nabla J_1(Su)(s)\, ds\,, \tag{4.24}$$

we thus obtain

$$\Big\langle \nabla J_1(Su), \tilde{S}h \Big\rangle = \langle p, Bh \rangle = \big\langle B^T p, h \big\rangle\ . \tag{4.25}$$

The function p in (4.24) is called the **adjoint** or the **adjoint state** for the system (4.20). It solves on $[0, T]$ the "backward" initial value problem (the **adjoint system**)

$$\dot{p} = -A^T p - \nabla J_1(Su)\,, \quad p(T) = 0\,, \tag{4.26}$$

as one sees when one computes its time derivative from (4.24). Thus,

$$\langle \nabla j(u), h \rangle = \Big\langle \nabla J_1(Su), \tilde{S}h \Big\rangle + \langle \nabla J_2(u), h \rangle = \big\langle B^T p, h \big\rangle + \langle \nabla J_2(u), h \rangle\ . \tag{4.27}$$

This means in particular that, if one wants to compute $j'(u; h) = \langle \nabla j(u), h \rangle$ for different directions h, one does not have to compute $\tilde{S}h$ for each h. It suffices to solve the adjoint system **once**.

Optimality conditions. We return to Problem 4.3. There we want to minimize

$$j(u) = J_1(Su) + J_2(u)\,.$$

Proposition 4.6 *Let u be a minimizer for Problem 4.3, assume that J_1 and J_2 are locally Lipschitz continuous and directionally differentiable. Then*

$$J_1'(Su; \tilde{S}h) + J_2'(u; h) \geq 0\,, \quad \textit{for all } h \in K(u). \tag{4.28}$$

Proof: This is a direct consequence of Proposition 4.1, since

$$j'(u; h) = J_1'(Su; \tilde{S}h) + J_2'(u; h)$$

by the chain rule. □

Application to ordinary differential equations. As before we consider the linear system

$$\dot{y} = Ay + Bu\,, \quad y(0) = y_0\,. \tag{4.29}$$

with the control constraint

$$u \in K = \{v \in L^2(0,T) : v(t) \in U_{ad} \text{ for a.a. } t \in (0,T)\}\,, \tag{4.30}$$

where $U_{ad} \subset \mathbb{R}^m$ is convex, closed and nonempty. We want to minimize

$$j(u) = J_1(Su) + J_2(u)\,. \tag{4.31}$$

Proposition 4.7 *Let u be a minimizer of (4.29) – (4.31), assume that $J_1, J_2 : L^2(0,T) \to \mathbb{R}$ have gradients $\nabla J_1(Su)$ and $\nabla J_2(u)$. Then we have*

$$\left\langle B^T p + \nabla J_2(u), w - u \right\rangle_{L^2} \geq 0\,, \quad \textit{for all } w \in K, \tag{4.32}$$

where p is the solution of the adjoint system

$$\dot{p} = -A^T p - \nabla J_1(Su)\,, \quad p(T) = 0\,. \tag{4.33}$$

Proof: We have

$$0 \leq \langle \nabla j(u), h \rangle = \left\langle B^T p, h \right\rangle + \langle \nabla J_2(u), h \rangle\ , \quad \text{for all } h \in K(u),$$

according to Proposition 4.6 and (4.27). This is equivalent to (4.32), since $h \in K(u)$ if and only if $\lambda h = w - u$ for some $w \in K$ and some $\lambda > 0$. □

We thus obtain the **optimality system**

$$\begin{aligned} \dot{y} = Ay + Bu\,, &\quad y(0) = y_0\,, \\ \dot{p} = -A^T p - \nabla J_1(Su)\,, &\quad p(T) = 0\,, \\ u \in K\,, \quad \left\langle B^T p + \nabla J_2(u), w - u \right\rangle_{L^2} &\geq 0\,, \quad \text{for all } w \in K. \end{aligned} \tag{4.34}$$

It consists of ordinary differential equations and a variational inequality, and it has to be solved for the unknown functions y, p and u.

Let us consider the special case

$$J_2(u) = \frac{\alpha}{2} \int_0^T \|u(t)\|^2\, dt\,. \tag{4.35}$$

Then $\nabla J_2(u) = \alpha u$, and the variational inequality becomes

$$u \in K\,, \quad \left\langle B^T p + \alpha u, w - u \right\rangle_{L^2} \geq 0\,, \quad \text{for all } w \in K.$$

This is equivalent to

$$u \in K\,, \quad \left\langle -\frac{1}{\alpha} B^T p - u, w - u \right\rangle_{L^2} \leq 0\,, \quad \text{for all } w \in K.$$

This in turn is equivalent to

$$u = P_K\big(-\frac{1}{\alpha}B^T p\big)\,, \tag{4.36}$$

where P_K is the projection onto K in $L^2(0,T)$. Thus, in this case the variational inequality can be replaced by an equation.

So far, the differential equations in (4.34) are formulated pointwise in time, whereas the variational inequality

$$u \in K\,, \quad \int_0^T \big\langle B^T p(t) + \nabla J_2(u)(t), w(t) - u(t)\big\rangle\, dt \geq 0\,, \quad \text{for all } w \in K, \tag{4.37}$$

is not. Again, as it was sketched in Chapter 2, one can pass from (4.37) to a pointwise formulation

$$u(t) \in U_{ad}\,, \quad \big\langle B^T p(t) + \nabla J_2(u)(t), \omega - u(t)\big\rangle \geq 0\,, \quad \text{for all } \omega \in U_{ad}, \tag{4.38}$$

which holds for almost all $t \in (0,T)$. In the special case (4.35) one obtains analogously that the minimizer u has to satisfy

$$u(t) = P_{U_{ad}}\big(-\frac{1}{\alpha}B^T p(t)\big)\,, \quad \text{for a.a. } t \in (0,T). \tag{4.39}$$

Here, $P_{U_{ad}}$ is the projection onto U_{ad} in $\mathbb{R}^m$. The function

$$S(t) = -\frac{1}{\alpha}B^T p(t) \tag{4.40}$$

takes on a role similar to that of a switching function. The optimal control u in general is not bang-bang, since

$$u(t) = -\frac{1}{\alpha}B^T p(t) \in \operatorname{int}(U_{ad}) \tag{4.41}$$

whenever the right-hand side lies in $\operatorname{int}(U_{ad})$. In particular, (4.41) holds for all t if $U_{ad} = \mathbb{R}^m$ (the case where there is no control constraint).

Interpretation of the adjoint system as a Lagrange multiplier. We write the initial value problem as an equality constraint in function space

$$F(y,u) = 0\,, \tag{4.42}$$

where $F = (F_1, F_2)$ and

$$\begin{aligned} F_1(y,u)(t) &= \dot{y}(t) - Ay(t) - Bu(t)\,, \quad t \in (0,T)\,, \\ F_2(y,u) &= y(0) - y_0\,. \end{aligned} \tag{4.43}$$

For the moment we ignore how the function spaces are chosen. We define the Lagrange function

$$\begin{aligned} L(y,u,p,r) &= J_1(y) + J_2(u) - \langle p, F_1(y,u)\rangle_{L_2} - \langle r, F_2(y,u)\rangle_{\mathbb{R}^n} \\ &= J_1(y) + J_2(u) - \langle p, \dot{y} - Ay - Bu\rangle_{L_2} - \langle r, y(0) - y_0\rangle_{\mathbb{R}^n}\,. \end{aligned} \tag{4.44}$$

The dependence on p and r is linear. Thus,

$$\begin{aligned} \nabla_p L(y,u,p,r) &= \dot{y} - Ay - Bu\,, \\ \nabla_r L(y,u,p,r) &= y(0) - y_0\,. \end{aligned} \tag{4.45}$$

This means that $y = Su$ is the solution of the original system if and only if

$$\nabla_p L(y,u,p,r) = 0 = \nabla_r L(y,u,p,r)\,. \tag{4.46}$$

For the partial derivative with respect to y we obtain

$$\langle \nabla_y L(y,u,p,r), z\rangle = \langle \nabla J_1(y,z\rangle - \langle p, \dot{z} - Az\rangle - \langle r, z(0)\rangle\ . \tag{4.47}$$

Assuming that p is differentiable and partial integration is valid, we get

$$\begin{aligned} \langle \nabla_y L(y,u,p,r), z\rangle = \int_0^T \big\langle (\nabla J_1(y))(t) + \dot{p}(t) + A^T p(t), z(t)\big\rangle\ dt \\ - \langle p(T), z(T)\rangle + \langle p(0), z(0)\rangle - \langle r, z(0)\rangle\ . \end{aligned} \tag{4.48}$$

If p solves the adjoint system (4.33) and if $r = p(0)$, then the right-hand side of (4.48) is zero for "all" functions z, so

$$\nabla_y L(y,u,p,r) = 0\,. \tag{4.49}$$

The partial derivative with respect to u becomes

$$\begin{aligned} \langle \nabla_u L(y,u,p,r), h\rangle &= \langle \nabla J_2(u,h\rangle - \langle p, -Bh\rangle = \big\langle B^T p + \nabla J_2(u,h\big\rangle \\ &= \langle \nabla j(u), h\rangle\ . \end{aligned} \tag{4.50}$$

If the control u is optimal, then $\langle \nabla j(u), h\rangle \geq 0$ for all admissible directions h at u.

Summarizing, we obtain ("$(*)$" stands for (y,u,p,r))

$$\begin{aligned} \nabla_y L(*) = \nabla_p L(*) = \nabla_r L(*) = 0\,, \\ \langle \nabla_u L(*), h\rangle \geq 0 \qquad \text{for all } h \in K(u). \end{aligned} \tag{4.51}$$

To make these computations precise, one needs a function space for y and p such that $\dot{y}, \dot{p} \in L^2(0,T)$. This is achieved by the Sobolev space $H^1(0,T)$, which is a Hilbert space. Moreover, in $H^1(0,T)$ the rule of partial integration is valid.

Alternatively, one can stay within the framework of L^2 if one rewrites the initial value problem as an integral equation. Set

$$G(y,u)(t) = y(t) - y_0 - \int_0^t Ay(s) + Bu(s)\,ds\,, \quad t \in (0,T)\,. \tag{4.52}$$

The constraint then becomes

$$G(y,u) = 0\,, \quad G : L^2(0,T) \times L^2(0,T) \to L^2(0,T)\,. \tag{4.53}$$

We define the Lagrange function

$$L(y,u,q) = J_1(y) + J_2(u) - \langle q, G(y,u)\rangle\ , \quad q \in L^2(0,T)\,. \tag{4.54}$$

Then

$$\nabla_q L(y,u,q) = 0 \qquad \Leftrightarrow \qquad G(y,u) = 0\,. \tag{4.55}$$

Now

$$\langle \nabla_y L(y,u,q), z\rangle = \int_0^T \langle (\nabla J_1(y))(t), z(t)\rangle \; dt - \int_0^T \left\langle q(t), z(t) - \int_0^t Az(s)\,ds \right\rangle \, dt\,. \tag{4.56}$$

We define

$$p(t) = \int_t^T q(s)\,ds\,. \tag{4.57}$$

Using partial integration, a similar computation as above yields that if p solves the adjoint system (4.33), then

$$\begin{aligned} \nabla_y L(y,u,q) &= 0\,, \\ \langle \nabla_u L(y,u,q), h\rangle &\geq 0 \qquad \text{for all } h \in K(u). \end{aligned} \tag{4.58}$$

5 The Superposition Operator

In this section we denote by $|\cdot|$ a norm in $\mathbb{R}^n$.

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be given. We consider the superposition operator

$$(Fu)(x) = f(u(x)), \quad x \in \Omega, \tag{5.1}$$

which maps functions $u : \Omega \to \mathbb{R}^n$ to $Fu : \Omega \to \mathbb{R}^m$; here, Ω is assumed to be an open bounded subset of $\mathbb{R}^l$. We consider the situations

$$F : L^\infty(\Omega;\mathbb{R}^n) \to L^\infty(\Omega;\mathbb{R}^m), \quad F : L^2(\Omega;\mathbb{R}^n) \to L^1(\Omega;\mathbb{R}^m). \tag{5.2}$$

In order that F maps L^∞ into L^∞, it suffices that f is measurable and bounded on bounded arguments, that is,

$$\sup_{|v|\le M} |f(v)| < \infty, \quad \text{for all } M > 0. \tag{5.3}$$

This is the case, for example, if f is continuous. In order that F maps L^2 into L^1, it suffices that f is measurable and satisfies, for some constants a and b,

$$|f(v)| \le a + b|v|^2, \quad \text{for all } v \in \mathbb{R}^n. \tag{5.4}$$

Then

$$\|Fu\|_{L^1} = \int_\Omega |f(u(x))|\,dx \le \int_\Omega a + b|u(x)|^2\,dx = a|\Omega| + b\|u\|_{L^2}^2 < \infty, \tag{5.5}$$

if $u \in L^2$. It is a nontrivial result (which we do not need) that this condition is also necessary, that is, if (5.4) does not hold for some a and b, then F does not map L^2 into L^1.

We want to investigate whether F is differentiable. If f is continuously differentiable, we have

$$f(u(x) + v) - f(u(x)) = Df(u(x))v + o(|v|).$$

The natural candidate for the derivative of F at u in the direction h is therefore

$$(DF(u)h)(x) = Df(u(x))h(x), \quad x \in \Omega. \tag{5.6}$$

In the case of L^∞, the mapping $h \mapsto Df(u(\cdot))h(\cdot)$ indeed defines a linear and continuous map from L^∞ to L^∞, since Df is continuous and therefore the function $x \mapsto Df(u(x))$ is bounded. In the case of L^2, we additionally assume that

$$\|Df(v)\| \le a + b|v|, \quad \text{for all } v \in \mathbb{R}^n. \tag{5.7}$$

Then $x \mapsto Df(u(x))$ belongs to L^2 whenever $u \in L^2$, and thus the right hand side of (5.6) belongs to L^1 for $h \in L^2$, by the Cauchy-Schwarz inequality.

In order to estimate the remainder term, we consider the mapping

$$g_x(t) = f(u(x) + th(x)), \quad g : (-\delta, \delta) \to \mathbb{R}^m. \tag{5.8}$$

We then have, since g is continuously differentiable,

$$\begin{aligned}(F(u+h))(x) - (Fu)(x) = g_x(1) - g_x(0) &= \int_0^1 g_x'(t)\,dt \\ &= \int_0^1 Df(u(x)+th(x))h(x)\,dt\,.\end{aligned} \tag{5.9}$$

We consider the remainder term $r_h : \Omega \to \mathbb{R}^m$ defined by

$$\begin{aligned}r_h(x) &= (F(u+h))(x) - (Fu)(x) - Df(u(x))h(x) \\ &= \int_0^1 Df(u(x)+th(x)) - Df(u(x))\,dt \cdot h(x)\,.\end{aligned} \tag{5.10}$$

We estimate it pointwise by

$$|r_h(x)| \le \int_0^1 \|Df(u(x)+th(x)) - Df(u(x))\|\,dt \cdot |h(x)|\,, \quad x \in \Omega\,. \tag{5.11}$$

In order to proceed further, we want to estimate the integral if h is close to 0. In the case of L^∞ this means that almost all values $u(x)$ and $h(x)$ are bounded resp. close to 0. In this case, it suffices that $Df(u)$ satisfies in every bounded set B a Lipschitz condition

$$\|Df(v) = Df(\tilde{v}\| \le L|v-\tilde{v}|\,, \quad \text{for all } v, \tilde{v} \in B. \tag{5.12}$$

This is equivalent to Df being locally Lipschitz continuous. In the case of L^2 the values of u and h may be unbounded no matter how small the norm of u and h is. In this case we require Df to be globally Lipschitz continuous, that is, (5.12) holds for $B = \mathbb{R}^n$. The estimate (5.11) then becomes in both cases

$$|r_h(x)| \le L|h(x)|^2\,. \tag{5.13}$$

In the case of L^∞ we then get

$$\|r_h\|_\infty \le L\|h\|_\infty^2\,, \quad \lim_{\|h\|_\infty \to 0} \frac{\|r_h\|_\infty}{\|h\|_\infty} = 0\,. \tag{5.14}$$

In the case of L^2 we get

$$\|r_h\|_1 \le L\int_\Omega |h(x)|^2\,dx = \|h\|_2^2\,, \quad \lim_{\|h\|_2 \to 0} \frac{\|r_h\|_1}{\|h\|_2} = 0\,. \tag{5.15}$$

Thus, in both cases F is differentiable.

In order to investigate whether the mapping $u \mapsto DF(u)$ is continuous, we consider the expression

$$\|DF(u) - DF(\tilde{u})\| = \sup_{\|h\|=1} \|(DF(u) - DF(\tilde{u}))h\|\,. \tag{5.16}$$

We have

$$|((DF(u) - DF(\tilde{u}))h)(x)| = |(Df(u(x)) - Df(\tilde{u}(x)))(h(x))| \le L|u(x) - \tilde{u}(x)|\,|h(x)|\,. \tag{5.17}$$

Thus, in the case of L^∞,

$$\begin{aligned} \|DF(u) - DF(\tilde{u})\| &= \sup_{\|h\|_\infty=1} \|(DF(u) - DF(\tilde{u}))h\|_\infty \\ &\leq \sup_{\|h\|_\infty=1} L\|u-\tilde{u}\|_\infty \|h\|_\infty = L\|u-\tilde{u}\|_\infty \,. \end{aligned} \tag{5.18}$$

In the case of L^2,

$$\begin{aligned} \|DF(u) - DF(\tilde{u})\| &= \sup_{\|h\|_2=1} \|(DF(u) - DF(\tilde{u}))h\|_1 \\ &\leq \sup_{\|h\|_2=1} \int_\Omega |(Df(u(x)) - Df(\tilde{u}(x)))(h(x))|\, dx \\ &\leq \sup_{\|h\|_2=1} \int_\Omega L|u(x)-\tilde{u}(x)|\,|h(x)|\, dx \leq \sup_{\|h\|_2=1} L\|u-\tilde{u}\|_2 \|h\|_2 \\ &= L\|u-\tilde{u}\|_2 \,. \end{aligned} \tag{5.19}$$

We summarize the above results for the superposition operator defined by

$$(Fu)(x) = f(u(x)) \,. \tag{5.20}$$

Proposition 5.1 *Assume that $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable and that*
(i) Df is locally Lipschitz continuous, or
(i') f satisfies (5.4) and Df satisfies (5.7) for some $a, b > 0$, and Df is globally Lipschitz continuous.
Then F defined by (5.20) is continuously differentiable from L^∞ to L^∞ in case (i), and from L^2 to L^1 in case (i'). The derivative of F is given by

$$(DF(u)h)(x) = Df(u(x))h(x) \,, \quad x \in \Omega \,. \tag{5.21}$$

□

More information on superposition operators can be found in the book of J. Appell and P. Zabrejko with the title "Nonlinear superposition operators".

If f satisfies the growth condition

$$|f(v)| \leq a + b|v| \,, \tag{5.22}$$

then F maps L^2 into L^2. But one can prove that $F : L^2 \to L^2$ is not differentiable unless f is affine. Consider the special case $n = m = 1$, $\Omega = (0,1)$. As shown above, the remainder can be written as

$$r_h(x) = \int_0^1 f'(u(x) + th(x)) - f'(u(x))\, dt \cdot h(x) \,. \tag{5.23}$$

Let u be a constant function, set

$$h_\varepsilon = 1_{(0,\varepsilon)} \,, \quad h_\varepsilon(x) = \begin{cases} 1 \,, & x < \varepsilon \,, \\ 0 \,, & x \geq \varepsilon \,. \end{cases} \tag{5.24}$$

Then

$$r_\varepsilon(x) = \int_0^1 f'(u+t) - f'(u)\,dt \cdot h_\varepsilon(x)\,. \tag{5.25}$$

Setting

$$d = \int_0^1 f'(u+t) - f'(u)\,dt\,,$$

we obtain, if $d \neq 0$, $\|r_\varepsilon\|_2 = |d|\|h_\varepsilon\|_2$, thus

$$\frac{\|r_\varepsilon\|_2}{\|h_\varepsilon\|_2} = |d| \neq 0\,, \quad \|h_\varepsilon\|_2 = \sqrt{\varepsilon} \to 0 \quad \text{as } \varepsilon \to 0. \tag{5.26}$$

Therefore, $F : L^2 \to L^2$ is not differentiable at such a point u.

6 Second Order Conditions

Second order derivatives. Let X, Y be normed spaces, $F : X \to Y$. For $u \in X$, we define the **second order directional derivative** of F in the direction (h, k) by

$$F''(u; h, k) = \lim_{\lambda \downarrow 0} \frac{F'(u + \lambda k; h) - F'(u; h)}{\lambda}\,, \tag{6.1}$$

if the limit exists, and if F is directionally differentiable in a neighbourhood of u (thus, $F'(u + \lambda k; h)$ exists for small enough λ.) If this is the case for all $h, k \in X$, we say that F is **twice directionally differentiable** at u.

If the mapping $(h, k) \to F''(u; h, k)$ is bilinear and continuous, we denote it by

$$D^2F(u) : X \times X \to Y\,, \quad D^2F(u)(h, k) = F''(u; h, k)\,. \tag{6.2}$$

In the special case $X = \mathbb{R}^n$, $Y = \mathbb{R}$, $D^2F(u)$ can be identified with the Hessian $H_F(u) \in \mathbb{R}^{(n,n)}$,

$$D^2F(u)(h, k) = h^T H_F(u) h\,. \tag{6.3}$$

The set of bilinear continuous mappings from X to Y is a normed space with the norm

$$\|T\| = \sup_{\|h\|_X = 1} \sup_{\|k\|_X = 1} \|T(h, k)\|_Y\,, \tag{6.4}$$

and we have

$$\|T(h, k)\|_Y \le \|T\| \|h\|_X \|k\|_Y\,, \quad \text{for all } h, k \in X. \tag{6.5}$$

One usually computes $D^2F(u)$ by evaluating (6.1).

If moreover the mapping $u \mapsto D^2F(u)$ is continuous on some open set $V \subset X$, then F is called **twice continuously differentiable** on V. As in the finite-dimensional case, Taylor's formula then holds, namely

$$F(u + h) = F(u) + DF(u)h + \frac{1}{2} D^2F(u)(h, h) + r_2(h)\,, \quad \lim_{h \to 0} \frac{r_2(h)}{\|h\|_X^2} = 0\,. \tag{6.6}$$

Second order optimality conditions. We again consider the problem

$$\text{minimize } j(u), \quad u \in K\,, \tag{6.7}$$

where X is a normed space, $j : X \to \mathbb{R}$ and $K \subset X$ is convex. If u is a minimizer, we know that $j'(u; h) \ge 0$ for all admissible directions, that is, for all $h \in K(u)$. This is a necessary condition for optimality. The next proposition provides a sufficient condition in terms of second derivatives.

Proposition 6.1 *Let $j : X \to \mathbb{R}$ be twice continuously differentiable, let $u \in K$ such that $Dj(u)h \ge 0$ for all $h \in K(u)$. Assume that there exists a $\gamma > 0$ such that*

$$D^2j(u)(h, h) \ge \gamma \|h\|^2\,, \quad \textit{for all } h \in K(u). \tag{6.8}$$

Then u is a strict local minimizer, and there exists a $\delta > 0$ such that

$$j(v) \ge j(u) + \frac{\gamma}{4} \|v - u\|^2 \tag{6.9}$$

holds for all $v \in K$ with $\|v - u\| < \delta$.

Proof: Let $v \in K$ be arbitrary. Then $h = v - u \in K(u)$ and

$$j(v) = j(u+h) = j(u) + Dj(u)h + \frac{1}{2}D^2j(u)(h,h) + r(h)\,, \quad \lim_{h\to 0}\frac{r(h)}{\|h\|^2} = 0\,.$$

Due to the assumptions,

$$j(u+h) \geq j(u) + \frac{\gamma}{2}\|h\|^2 + r(h)\,.$$

Let $\delta > 0$ such that $|r(h)| \leq (\gamma/4)\|h\|^2$ for every h with $\|h\| < \delta$. □

In the case $K = X$ (the case without a control constraint), we have $K(u) = X$, and the above proposition has a simpler form.

Corollary 6.2 *Let $j : X \to \mathbb{R}$ be twice continuously differentiable, let $u \in X$ such that $Dj(u) = 0$. Assume that there exists a $\gamma > 0$ such that*

$$D^2j(u)(h,h) \geq \gamma\|h\|^2\,, \quad \textit{for all } h \in X. \tag{6.10}$$

Then u is a strict local minimizer, and there exists a $\delta > 0$ such that

$$j(v) \geq j(u) + \frac{\gamma}{4}\|v-u\|^2 \tag{6.11}$$

holds for all $v \in X$ with $\|v - u\| < \delta$. □

Two-norm discrepancy. Concerning our minimization problem, we have the two requirements

$$\lim_{h\to 0}\frac{r(h)}{\|h\|^2} = 0\,, \qquad \frac{A(h,h)}{\|h\|^2} \geq \gamma \text{ for all } h \in X, \tag{6.12}$$

where the bilinear form A has to be equal to $D^2j(u)$ if the limit exists. If X has finite dimension, this is true either for every norm or for no norm (modulo the size of γ), because all norms on X are equivalent. If X is infinite dimensional, it may happen that

- for some norm, the limit exists but $A = D^2j(u)$ is not positive definite,
- if we weaken the norm, then A becomes positive definite, but the limit (and thus $D^2j(u)$) no longer exists.

We illustrate this situation with the following example. (See F. Tröltzsch, Optimal Control of Partial Differential Equations, 2010; German edition 2005 and 2009.) We want to minimize

$$j(u) = -\int_0^1 \cos(u(x))\,dx\,, \quad 0 \leq u(x) \leq 2\pi\,. \tag{6.13}$$

We consider the sets of solutions in the spaces $L^2(0,1)$ and $L^\infty(0,1)$. These sets are identical, namely they consist of the measurable functions with $u(x) = 1$, that is,

$$u(x) \in \{0, 2\pi\}\,, \quad \text{for all } x \in (0,1). \tag{6.14}$$

Let u_1 and u_2 be two different solutions. Then $|u_1(x) - u_2(x)| = 2\pi$ if $u_1(x) \neq u_2(x)$, and therefore

$$\|u_1 - u_2\|_\infty = 2\pi\,, \quad \|u_1 - u_2\|_2 = 2\pi\sqrt{\operatorname{meas}\{u_1 \neq u_2\}}\,. \tag{6.15}$$

Thus, every solution is a strict local minimum in L^∞, but no solution is a strict local minimum in L^2.

Now let us consider the second order optimality conditions. We have

$$j(u) = \int_0^1 (Fu)(x)\,dx\,, \tag{6.16}$$

where F is the superposition operator generated by $f(v) = -\cos v$. Since f belongs to C^∞ and all derivatives of f are bounded, the assumptions of Proposition 5.1 are satisfied, so $F : L^\infty \to L^\infty$ as well as $F : L^2 \to L^1$ are continuously differentiable with

$$(DF(u)h)(x) = f'(u(x))h(x)\,, \quad x \in (0,1)\,. \tag{6.17}$$

Since the integral is linear and continuous on L^1 (and therefore, on L^∞ too), the chain rule yields

$$Dj(u)h = \int_0^1 f'(u(x))h(x)\,dx \tag{6.18}$$

in both cases, $j : L^\infty \to \mathbb{R}$ and $j : L^2 \to \mathbb{R}$.

Now let us consider the second derivative in L^∞. For a given $h \in L^\infty$, the mapping $u \mapsto Dj(u)h$ can be expressed as the composition of the superposition operator

$$(Gu)(x) = f'(u(x))\,, \quad G : L^\infty \to L^\infty\,, \tag{6.19}$$

and the linear continuous mapping from L^∞ to $\mathbb{R}$

$$d \mapsto \int_0^1 d(x)h(x)\,dx\,. \tag{6.20}$$

We have $(G'(u)k)(x) = f''(u(x))k(x)$ and therefore

$$j''(u;(h,k)) = \int_0^1 f''(u(x))h(x)k(x)\,dx = \int_0^1 \cos(u(x))h(x)k(x)\,dx\,. \tag{6.21}$$

The mapping $(h,k) \mapsto j''(u;(h,k))$ is a bilinear and continuous mapping from $L^\infty \times L^\infty$ to $\mathbb{R}$, so

$$D^2j(u)(h,k) = \int_0^1 f''(u(x))h(x)k(x)\,dx = \int_0^1 \cos(u(x))h(x)k(x)\,dx\,. \tag{6.22}$$

Moreover, $u \mapsto D^2j(u)$ is continuous, so j is twice continuously differentiable on L^∞. However, if we set

$$h_\varepsilon = 1_{(0,\varepsilon)}$$

we have $\|h_\varepsilon\|_\infty = 1$ and

$$\frac{D^2j(u)(h_\varepsilon,h_\varepsilon)}{\|h_\varepsilon\|_\infty^2} \leq \int_0^1 h_\varepsilon(x)^2\,dx = \varepsilon\,. \tag{6.23}$$

Therefore $D^2 j(u)$ is not positive definite on L^∞, and (6.9) does not hold in L^∞, no matter how we choose u. In particular, at this point we cannot use the second derivative in order to conclude that $u = 0$ is a strict local minimum in L^∞ (which is true according to (6.15)).

On the other hand, the quadratic form

$$A_u(h,k) = \int_0^1 \cos(u(x))h(x)k(x)\,dx \tag{6.24}$$

is well-defined on $L^2 \times L^2$, and for the minimizer $u = 0$ we obtain

$$A_0(h,h) = \int_0^1 h(x)^2\,dx = \|h\|_2^2\,, \quad \frac{A_0(h,h)}{\|h\|_2^2} = 1\,, \tag{6.25}$$

so A is positive definite on L^2. However, it turns out that if we regard j as a functional from L^2 to $\mathbb{R}$, the second derivative $D^2 j(u)$ does not exist in $u = 0$. Indeed, setting

$$g_x(t) = f(u(x) + th(x)) = f(th(x))\,,$$

we have the Taylor expansion

$$g_x(1) = g_x(0) + g_x'(0) + \frac{1}{2}g_x''(0) + \int_0^1 (g_x''(t) - g_x''(0))(1-t)\,dt\,,$$

so

$$f(h(x)) = f(0) + f'(0)h(x) + \frac{1}{2}f''(0)h(x)^2 + \int_0^1 (1-t)(f''(th(x)) - f''(0))h(x)^2\,dt\,.$$

We integrate over $(0,1)$ with respect to x and obtain

$$j(h) = j(0) + Dj(0)h + \frac{1}{2}\int_0^1 f''(0)h(x)^2\,dx + r_2(h)\,, \tag{6.26}$$

where

$$r_2(h) = \int_0^1 \int_0^1 (1-t)(f''(th(x)) - f''(0))\,h(x)^2\,dt\,dx\,. \tag{6.27}$$

Again we choose

$$h_\varepsilon = 1_{(0,\varepsilon)}$$

and get

$$r_2(h_\varepsilon) = \frac{1}{2}\int_0^\varepsilon \int_0^1 (1-t)(\cos t - \cos 0)\,dt\,dx = \varepsilon c \tag{6.28}$$

where $c \neq 0$. Since $\|h_\varepsilon\|_2 = \sqrt{\varepsilon}$, we finally obtain

$$\frac{r_2(h_\varepsilon)}{\|h_\varepsilon\|_2^2} = c \neq 0\,. \tag{6.29}$$

Thus, j is not twice continuously differentiable at $u = 0$.

Again, we cannot use Proposition 6.1 to conclude optimality of $u = 0$ (and indeed, $u = 0$ is not a strict local minimizer in L^2).

However, one can combine the results concerning L^∞ and L^2. From (6.27) we get, since $|f'''| \le 1$,

$$|r_2(h)| \le \int_0^1 \int_0^1 (1-t)t\,dt \cdot |h(x)|^3\,dx \le \frac{1}{6}\|h\|_\infty \|h\|_2^2 . \tag{6.30}$$

This implies that, if we restrict ourselves to $h \in L^\infty$,

$$\frac{r_2(h)}{\|h\|_2^2} \to 0 \qquad \text{as} \quad \|h\|_\infty \to 0 . \tag{6.31}$$

Now (6.26) becomes, since $Dj(0) = 0$ and $f''(0) = 1$,

$$\begin{aligned} j(h) &= j(0) + Dj(0)h + \frac{1}{2}\int_0^1 f''(0)h(x)^2\,dx + r_2(h) \\ &= j(0) + \frac{1}{2}\|h\|_2^2 + r_2(h) = j(0) + \|h\|_2^2\Big(\frac{1}{2} + \frac{r_2(h)}{\|h\|_2^2}\Big) \end{aligned} \tag{6.32}$$

From (6.31) we conclude that

$$j(h) > 0 , \quad \text{if } \|h\|_\infty \text{ is sufficiently small.} \tag{6.33}$$

This proves that $u = 0$ is a strict local minimizer in L^∞.

In this manner one can resolve the difficulty (of proving optimality) which arises from the two-norm discrepancy. This technique is used in particular in optimal control problems for partial differential equations.

7 The Semismooth Newton Method

One way to solve an optimization problem is to find a solution of the first order optimality system.

The simplest case arises when the cost functional is quadratic and no constraint is present. We consider

$$\text{minimize} \quad j(u) = \frac{1}{2}\langle u, Gu\rangle - \langle f, u\rangle \ . \tag{7.1}$$

Here, $j : X \to \mathbb{R}$, X is a Hilbert space, $f \in X$, and $G : X \to X$ is a linear and continuous operator.

We then have

$$\begin{aligned} \langle \nabla j(u), h\rangle &= \frac{1}{2}\langle h, Gu\rangle + \frac{1}{2}\langle u, Gh\rangle - \langle f, h\rangle \\ &= \left\langle \frac{1}{2}(G + G^*)u, h \right\rangle - \langle f, h\rangle \ , \end{aligned} \tag{7.2}$$

where $G^* : X \to X$ is the Hilbert adjoint of G, defined by

$$\langle G^* u, v\rangle = \langle u, Gv\rangle \ , \quad \text{for all } u, v \in X.$$

In fact, we may assume that G is self-adjoint, that is, $G^* = G$, since otherwise we may replace G with the self-adjoint operator $(G + G^*)/2$ without changing j. Then

$$\nabla j(u) = Gu - f \, . \tag{7.3}$$

All candidates for a minimizer of (7.1) must satisfy

$$Gu = f \, . \tag{7.4}$$

This is a linear equation in the space X. Since $D^2 j$ is constant and equal to G, solving (7.4) can be interpreted as performing a Newton step

$$D^2 j(u_0)(u - u_0) = -\nabla j(u_0)$$

for an arbitrary initial value $u_0 \in X$.

Active set strategy for inequality problems. We consider the linear-quadratic ODE control problem for the special case of a scalar control.

$$\text{Miminize } j(u) = J(Su, u), \quad J(y, u) = \frac{1}{2}\int_0^T \|y(t) - y_d(t)\|^2 \, dt + \frac{\alpha}{2}\int_0^T u(t)^2 \, dt \, , \tag{7.5}$$

where

$$\dot{y} = Ay + bu \, , \quad y(0) = y_0 \, , \tag{7.6}$$

with $A \in \mathbb{R}^{(n,n)}$, $b \in \mathbb{R}^n$, and

$$u \in K = \{v \in L^2(0, T) : v(t) \in [u_{min}, u_{max}] \text{ a.e. in } (0, T)\} \, . \tag{7.7}$$

The optimality system is given by (see Section 4)

$$\begin{aligned} \dot{y} &= Ay + bu\,, \quad y(0) = y_0\,, \\ \dot{p} &= -A^T p - (y - y_d)\,, \quad p(T) = 0\,, \\ u \in K\,, &\quad \big\langle b^T p + \alpha u, w - u \big\rangle_{L^2} \geq 0\,, \quad \text{for all } w \in K. \end{aligned} \tag{7.8}$$

It consists of ordinary differential equations and a variational inequality, and it has to be solved for the unknown functions y, p and u.

We have already seen that in this particular case the variational inequality is equivalent to

$$u = P_K\big(-\frac{1}{\alpha} b^T p\big)\,, \tag{7.9}$$

where P_K is the projection onto K in $L^2(0,T)$.

The pointwise formulation of (7.9) is

$$u(t) = P_{[u_{min}, u_{max}]}\big(-\frac{1}{\alpha} b^T p(t)\big)\,, \quad \text{for a.a. } t \in (0,T). \tag{7.10}$$

Setting

$$s(t) = -\frac{1}{\alpha} b^T p(t)\,, \tag{7.11}$$

we can rewrite (7.10) as

$$u(t) = \begin{cases} u_{min}\,, & s(t) < u_{min}\,, \\ u_{max}\,, & s(t) > u_{max}\,, \\ s(t)\,, & \text{otherwise.} \end{cases} \tag{7.12}$$

For an arbitrary function $\sigma : [0,T] \to \mathbb{R}$ we define the **active sets**

$$\begin{aligned} A^{min}(\sigma) &= \{t : t \in [0,T]\,,\, \sigma(t) < u_{min}\}\,, \\ A^{max}(\sigma) &= \{t : t \in [0,T]\,,\, \sigma(t) > u_{max}\}\,, \end{aligned} \tag{7.13}$$

and the **inactive set**

$$I(\sigma) = [0,T] \setminus (A^{max}(\sigma) \cup A^{min}(\sigma))\,. \tag{7.14}$$

The idea of the active set strategy is the following:
Given a current iterate $(y_{n-1}, p_{n-1}, u_{n-1}, s_{n-1})$ of the optimality system, we update the active sets by

$$\begin{aligned} A_n^{min} &= A^{min}(s_{n-1})\,, \\ A_n^{max} &= A^{max}(s_{n-1})\,, \end{aligned} \tag{7.15}$$

Then we compute the new iterate (y_n, p_n, u_n, s_n) from the optimality system, where we replace the nonlinear equation (7.9) by

$$u_n(t) = \begin{cases} u_{min}\,, & t \in A_n^{min}\,, \\ u_{max}\,, & t \in A_n^{max}\,, \\ s_n(t)\,, & \text{otherwise.} \end{cases} \tag{7.16}$$

In fact, this second part of the iteration is a linear problem. Let us denote by 1_n^{min} and 1_n^{max} the characteristic functions of A_n^{min} and A_n^{max}. Then (y_n, p_n, u_n, s_n) is the solution of the linear system

$$\begin{aligned} \dot{y} &= Ay + bu\,, \quad y(0) = y_0 \\ \dot{p} &= -A^T p - (y - y_d)\,, \quad p(T) = 0 \\ s &= -\frac{1}{\alpha} b^T p \\ u - (1 - 1_n^{min} - 1_n^{max})s &= 1_n^{min} u_{min} + 1_n^{max} u_{max} \end{aligned} \tag{7.17}$$

for the unknown functions (y, p, u, s) as functions of time. The results from the previous iteration step enter this system only through the active sets A_n^{min} and A_n^{max} which were obtained by (7.15). In this way, the active set strategy replaces the nonlinear optimality system by a sequence of linear problems. It is also called **primal-dual active set strategy**, since the update of the active sets is based on the solution of the system for the "primal" variable y and the "dual" variable p.

The semismooth Newton method. This is a variant of the Newton method for solving a nonlinear equation

$$F(u) = 0\,, \quad F : X \to Z\,, \tag{7.18}$$

where X and Z are Banach spaces. The Newton method itself is an iterative method which replaces the nonlinear equation (7.18) by a sequence of linear problems. Starting from an initial value $u_0 \in X$, it is defined by

$$\begin{aligned} DF(u_k)h_k &= -F(u_k) \\ u_{k+1} &= u_k + h_k\,. \end{aligned} \tag{7.19}$$

Thus, the increment h_k is determined as the solution of a linear equation. In optimization, F corresponds to ∇j, DF corresponds to $D^2 j$. The iteration step makes sense if F is differentiable at u_k with a linear and continuous derivative $DF(u_k) : X \to Z$ which is bijective. Then the linear equation has a unique solution h_k for given $F(u_k)$, and the inverse $DF(u_k)^{-1}$ is continuous, as a consequence of the open mapping theorem in functional analysis. Thus, h_k depends continuously upon $F(u_k)$.

If such a derivative is not available, or if for some reasons one does not want to use it, one may replace the operator $DF(u)$ by some other operator $G(u)$ which is linear, continuous and bijective. The iteration (7.19) becomes

$$\begin{aligned} G(u_k)h_k &= -F(u_k) \\ u_{k+1} &= u_k + h_k\,. \end{aligned} \tag{7.20}$$

Under suitable assumptions, the Newton method converges locally with a quadratic convergence rate to a $u_* \in X$ with $F(u_*) = 0$. When replacing DF with G, one loses the quadratic convergence. One is, however, interested to keep superlinear convergence,

$$\lim_{k\to\infty} \frac{\|u_{k+1} - u_*\|}{\|u_k - u_*\|} = 0\,. \tag{7.21}$$

In order to see which properties of G would guarantee that, we observe that

$$\begin{aligned} u_{k+1} - u_* &= u_k + h_k - u_* = u_k - u_* - G(u_k)^{-1}F(u_k) \\ &= G(u_k)^{-1}\big(G(u_k)(u_k - u_*) - F(u_k) + F(u_*)\big)\,. \end{aligned}$$

It follows that

$$\|u_{k+1} - u_*\| \le \|G(u_k)^{-1}\| \cdot \|F(u_k) - F(u_*) - G(u_k)(u_k - u_*)\| . \tag{7.22}$$

This leads to the following definition.

Definition 7.1 (Newton derivative)
Let X, Z be Banach spaces, $U \subset X$ open, $F : U \to Z$. A mapping $G : U \to L(X, Z)$ is called a ***Newton derivative*** *of F in U if*

$$\lim_{h \to 0} \frac{\|F(u+h) - F(u) - G(u+h)h\|}{\|h\|} = 0 \tag{7.23}$$

holds for all $u \in U$. In this case, F is called ***Newton differentiable*** *in U.*

If G is a Newton derivative of F, the iteration (7.20) for solving $F(u) = 0$ is called the **semismooth Newton method**.

The Newton derivative may not be unique. For example, if F is continuously differentiable, then $G(u) = F'(u - h)$ as well as $G(u) = F'(u)$ define Newton derivatives of F.

Proposition 7.2 (Superlinear convergence)
Let X, Z be Banach spaces, $U \subset X$ open, $F : U \to Z$, let $u_* \in U$ with $F(u_*) = 0$. Let G be a Newton derivative of F in U and assume that the set $\{\|G(u)^{-1}\| : u \in U\}$ is bounded. Then there exists an $\varepsilon > 0$ such that, for all initial values $u_0 \in B(u_*, \varepsilon)$, the iteration (7.20) is well-defined, and $u_k \to u_*$ superlinearly.

Proof: Let $\|G(u)^{-1}\| \le M$ in U. Choose an arbitrary $\eta \in (0, 1)$, and choose $\varepsilon > 0$ such that

$$\|F(u_* + h) - F(u_*) - G(u_* + h)h\| \le \frac{\eta}{M}\|h\| , \quad \text{for all } \|h\| < \varepsilon. \tag{7.24}$$

Choose any u_0 with $\|u_0 - u_*\| < \varepsilon$. Setting $h = u_0 - u_*$, we obtain from (7.22) that

$$\begin{aligned} \|u_1 - u_*\| &\le \|G(u_0)^{-1}\| \cdot \|F(u_0) - F(u_*) - G(u_0)(u_0 - u_*)\| \\ &\le M\frac{\eta}{M}\|u_0 - u_*\| = \eta\|u_0 - u_*\| , \end{aligned} \tag{7.25}$$

so $u_1 \in B(u_*, \varepsilon)$. Replacing u_0 with u_{k-1} and u_1 with u_k, we see by induction that $u_k \in B(u_*, \varepsilon)$ for all k and

$$\|u_{k+1} - u_*\| \le \eta\|u_k - u_*\| .$$

Thus the iteration is well-defined, and $u_k \to u_*$ linearly. By (7.22),

$$\frac{\|u_{k+1} - u_*\|}{\|u_k - u_*\|} \le M\frac{\|F(u_k) - F(u_*) - G(u_k)(u_k - u_*)\|}{\|u_k - u_*\|} \to 0$$

as $k \to \infty$, according to the definition of the Newton derivative, since $u_k \to u_*$. □

The chain rule holds for Newton derivatives. If G_1 is a Newton derivative for F_1, and G_2 a Newton derivative for F_2, and if F_1 is locally Lipschitz continuous, then $G_2 \circ G_1$ is a Newton derivative for $F_2 \circ F_1$. Like the proof of Proposition 4.5, it is solely based on the definitions. (See the paper of Hintermüller and Kunisch in SIAM J. Opt. 20 (2009), 1133 – 1156.)

Example 7.3
1. Let $F : \mathbb{R} \to \mathbb{R}$, $F(x) = x^+ = \max\{x, 0\}$. Then

$$G(x) = \begin{cases} 1\,, & x > 0\,, \\ 0\,, & x < 0\,, \\ d\,, & x = 0\,, \end{cases} \tag{7.26}$$

with $d \in \mathbb{R}$ arbitrary, is a Newton derivative of F in $\mathbb{R}$. Indeed, in $\{x \neq 0\}$ it coincides with the classical derivative for h small enough, and in $x = 0$ we have for all $h \in \mathbb{R}$

$$F(x + h) - F(x) - G(x + h)h = h^+ - 0 - h^+ = 0\,.$$

2. Since $|x| = x^+ + (-x)^+$, the function $F(x) = |x|$ has

$$G(x) = \begin{cases} 1\,, & x > 0\,, \\ -1\,, & x < 0\,, \\ d\,, & x = 0\,, \end{cases} \tag{7.27}$$

with $d \in \mathbb{R}$ arbitrary, as a Newton derivative.
3. The projection mapping $F : \mathbb{R} \to \mathbb{R}$ onto $[a, b]$,

$$F(x) = P_{[a,b]}(x) = \min\{b, \max\{x, a\}\} = \begin{cases} b\,, & x \geq b\,, \\ x\,, & a < x < b\,, \\ a\,, & x \leq a\,, \end{cases} \tag{7.28}$$

has as a Newton derivative the mapping

$$G(x) = \begin{cases} 1\,, & a < x < b\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{7.29}$$

This can be computed directly, or it can be reduced to $x \mapsto x^+$, since

$$P_{[a,b]}(x) = (b - a - (x - a)^+)^+ - b\,.$$

4. Let X be a Hilbert space, $F : X \to \mathbb{R}$, $F(x) = \|x\|$. Then $G : X \to X$ defined by (recall the identification of X^* and X)

$$G(x) = \frac{x}{\|x\|}\,, \quad x \neq 0\,, \tag{7.30}$$

and $G(0) = d$, $d \in X$ arbitrary, is a Newton derivative of F. This involves some scalar product computations, see the book of Ito and Kunisch.

The following result is due to Ulbrich, Hintermüller, Ito and Kunisch.

Proposition 7.4 *Let $\Omega \subset \mathbb{R}^n$ be open and bounded, $1 \leq p < q \leq \infty$, $F : L^q(\Omega) \to L^p(\Omega$ be defined by*

$$(Fu)(x) = u^+(x) = \max\{u(x), 0\}\,, \quad x \in \Omega\,. \tag{7.31}$$

Then $G : L^q(\Omega) \to L(L^q(\Omega), L^p(\Omega))$ defined by

$$((Gu)h)(x) = \begin{cases} h(x)\,, & u(x) > 0\,, \\ 0\,, & u(x) < 0\,, \\ d\,, & u(x) = 0\,, \end{cases} \tag{7.32}$$

with $d \in \mathbb{R}$ arbitrary, is a Newton derivative for F.

The assumption $p < q$ is essential, in the case $p = q$ the assertion is false. This is analogous to the situation for general superposition operators discussed in Section 5.

Corollary 7.5 *Let $\Omega \subset \mathbb{R}^n$ be open and bounded, $1 \le p < q \le \infty$, $F : L^q(\Omega) \to L^p(\Omega)$ be defined by*

$$(F_P u)(x) = P_{[a,b]}(u(x)) = \begin{cases} b\,, & u(x) \ge b\,, \\ u(x)\,, & a < u(x) < b\,, \\ a\,, & u(x) \le a\,. \end{cases} \tag{7.33}$$

Then $G_P : L^q(\Omega) \to L(L^q(\Omega), L^p(\Omega))$ defined by

$$((G_P u)h)(x) = \begin{cases} h(x)\,, & a < u(x) < b\,, \\ 0\,, & \textit{otherwise}\,, \end{cases} \tag{7.34}$$

is a Newton derivative of F_P.

Application to the control problem. We return to the optimal control problem with scalar control

$$\text{miminize } j(u) = J(Su, u), \quad J(y, u) = \frac{1}{2}\int_0^T \|y(t) - y_d(t)\|^2\,dt + \frac{\alpha}{2}\int_0^T u(t)^2\,dt\,, \tag{7.35}$$

where

$$\dot{y} = Ay + bu\,, \quad y(0) = y_0\,, \tag{7.36}$$

with $A \in \mathbb{R}^{(n,n)}$, $b \in \mathbb{R}^n$, and

$$u \in K = \{v \in L^2(0,T) : v(t) \in U_{ad} \text{ a.e. in } (0,T)\}\,, \quad U_{ad} = [u_{min}, u_{max}]\,. \tag{7.37}$$

As we have already computed,

$$\begin{aligned} \langle \nabla j(u), h\rangle &= \left\langle \nabla_y J(Su, u), \tilde{S}h \right\rangle + \langle \nabla_u J(Su, u), h\rangle \\ &= \left\langle Su - y_d, \tilde{S}h \right\rangle + \langle \alpha u, h\rangle\ . \end{aligned} \tag{7.38}$$

Here, $\tilde{S} : X \to Y$ is the linear part of S defined by $Su = \tilde{S}u + S0$, and $X = L^2(0,T)$, $Y = L^2(0,T;\mathbb{R}^n)$.

We rewrite (7.38) as

$$\langle \nabla j(u), h\rangle = \left\langle \tilde{S}^*(Su - y_d) + \alpha u, h \right\rangle\ . \tag{7.39}$$

Here, $\tilde{S}^* : Y \to X$ is the Hilbert adjoint of $\tilde{S} : X \to Y$, defined by

$$\left\langle \tilde{S}^* z, h \right\rangle = \left\langle z, \tilde{S} h \right\rangle , \quad z \in Y , \, h \in X . \tag{7.40}$$

We define $Q : X \to X$ by

$$Qu = \tilde{S}^*(Su - y_d) . \tag{7.41}$$

Then

$$\langle \nabla j(u), h \rangle = \langle Qu + \alpha u, h \rangle . \tag{7.42}$$

As before, the optimality condition for the minimizer u_*,

$$\langle \nabla j(u_*), v - u_* \rangle \geq 0 , \quad \text{for all } v \in K,$$

is equivalent to

$$u_* = P_K\big(-\frac{1}{\alpha} Q u_* \big) . \tag{7.43}$$

We want to solve this equation with the semismooth Newton method. To this end, we define

$$Fu = u - P_K\big(-\frac{1}{\alpha} Qu \big) , \quad F : X \to X . \tag{7.44}$$

Proposition 7.6 *The mapping $F : L^2(0,T) \to L^2(0,T)$ defined by (7.44) is Newton differentiable. A Newton derivative is given by*

$$(Gu)(h) = h + 1_{in} \cdot \frac{1}{\alpha} \tilde{S}^* \tilde{S} h , \tag{7.45}$$

where

$$1_{in}(t) = \begin{cases} 1 , & u_{min} < -\frac{1}{\alpha}(Qu)(t) < u_{max} , \\ 0 , & \textit{otherwise,} \end{cases} \tag{7.46}$$

is the characteristic function of the inactive set.

Proof: Since for K given by (7.37) the projection can be evaluated pointwise,

$$(P_K v)(t) = P_{U_{ad}}(v(t)) ,$$

we obtain from Corollary 7.5 that

$$P_K : L^q(0,T) \to L^2(0,T) \tag{7.47}$$

is Newton differentiable if $q > 2$. From the very definition, Q maps $X = L^2$ into itself, which is not good enough. However, in view of (4.25), we have

$$\langle Qu, h \rangle = \left\langle \tilde{S}^*(Su - y_d), h \right\rangle = \left\langle Su - y_d, \tilde{S} h \right\rangle = \left\langle b^T p, h \right\rangle ,$$

where p is the solution of the adjoint system

$$\dot{p} = -A^T p - (Su - y_d) , \quad p(T) = 0 . \tag{7.48}$$

The mapping $u \mapsto p$ defined by (7.48) is continuous and affine linear from $L^2(0,T)$ to $C[0,T]$, and the embedding $C[0,T] \to L^q(0,T)$ is continuous for all q. Therefore $Q : L^2(0,T) \to L^q(0,T)$ is continuous for $q > 2$. The derivative of Q, $Qu = \tilde{S}^*(Su - y_d)$, does not depend on u and is given by its linear part,

$$DQ(u) = \tilde{S}^*\tilde{S}\,. \tag{7.49}$$

According to Corollary 7.5, a Newton derivative of P_K at the point $s \in L^q(0,T)$ is given by

$$((G_P s)d)(t) = \begin{cases} d(t)\,, & u_{min} < s(t) < u_{max}\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

The chain rule now yields (7.46), setting

$$s = -\frac{1}{\alpha}Qu\,, \quad d = -\frac{1}{\alpha}DQ(u)(h) = -\frac{1}{\alpha}\tilde{S}^*\tilde{S}h\,.$$

□

Proposition 7.7 *The Newton derivative G given in Proposition 7.6 has an inverse which is uniformly bounded, that is, there exists M such that*

$$\|(Gu)^{-1}\| \le M$$

for all $u \in X$.

Proof: Let $u \in X$ be arbitary. According to Proposition 7.6, we have

$$(Gu)(h) = h + 1_{in} \cdot \frac{1}{\alpha}\tilde{S}^*\tilde{S}h\,, \tag{7.50}$$

where

$$1_{in}(t) = \begin{cases} 1\,, & u_{min} < -\frac{1}{\alpha}(Qu)(t) < u_{max}\,, \\ 0\,, & \text{otherwise.} \end{cases} \tag{7.51}$$

We also define

$$1_{act}(t) = 1 - 1_{in}(t)\,.$$

Let $h \in X$ be arbitrary, set $z = (Gu)h$. Then

$$h1_{act} = z1_{act}\,,$$

since $1_{in}1_{act} = 0$, and hence ("$\|\cdot\|$" denotes the norm in L^2)

$$\|h1_{act}\| = \|z1_{act}\| \le \|z\|\,. \tag{7.52}$$

In order to estimate $\|h1_{in}\|$, we test (7.50) with $h1_{in}$ and compute

$$\begin{aligned} \langle z, h1_{in}\rangle &= \langle h, h1_{in}\rangle + \frac{1}{\alpha}\left\langle \tilde{S}^*\tilde{S}h, h1_{in}\right\rangle \\ &= \langle h1_{in}, h1_{in}\rangle + \frac{1}{\alpha}\left\langle \tilde{S}^*\tilde{S}(h1_{in} + h1_{act}), h1_{in}\right\rangle \\ &= \|h1_{in}\|^2 + \frac{1}{\alpha}\left\langle \tilde{S}(h1_{in}), \tilde{S}(h1_{in})\right\rangle + \frac{1}{\alpha}\left\langle \tilde{S}(h1_{act}), \tilde{S}(h1_{in})\right\rangle\,. \end{aligned}$$

This yields

$$\begin{aligned}\|h1_{in}\|^2 &\le \langle z, h1_{in}\rangle - \frac{1}{\alpha}\left\langle \tilde{S}(h1_{act}), \tilde{S}(h1_{in})\right\rangle \le \|h1_{in}\|\big(\|z\| + \frac{1}{\alpha}\|\tilde{S}\|^2\|h1_{act}\|\big)\\ &\le \|h1_{in}\|C\|z\|\,, \quad C = 1 + \frac{1}{\alpha}\|\tilde{S}\|^2\,,\end{aligned}$$

where we have used (7.52) at the end. This yields

$$\|h\| \le \|h1_{in}\| + \|h1_{act}\| \le M\|z\|\,, \quad M = C + 1\,. \tag{7.53}$$

In particular, Gu is injective, since $\|z\| = 0$ implies $\|h\| = 0$. Next, the operator $\tilde{S}^*\tilde{S} : L^2 \to L^2$ is a compact operator. Thus, Gu has the form

$$Gu = I - K$$

where K is compact. The Riesz-Schauder theory from functional analysis then says that the Fredholm alternative holds for Gu; in particular, Gu is surjective if and only if Gu is injective. Therefore, Gu is bijective. The assertion now follows from (7.53). □

As a consequence, Proposition 7.2 can be applied. This shows that the semismooth Newton iteration locally converges to a minimizer of the control problem with a superlinear rate.

8 Bellman Principle and Dynamic Programming

Problem 8.1 (Discrete optimal control problem)
We consider a discrete evolution with initial value x, states $y_k \in G$ and controls $w_k \in W$,

$$y_{k+1} = g(y_k, w_k)\,, \quad y_0 = x \in G\,, \tag{8.1}$$

where $g : G \times W \to G$ and G, W are arbitrary sets. We define the set $\mathcal{W}$ of admissible controls by

$$\mathcal{W} = \{w : w = (w_0, w_1, \ldots), w_k \in W \text{ for all } k \in \mathbb{N}\}\,. \tag{8.2}$$

According to (8.1) we define by

$$\varphi(w; x) = (y_0, y_1, \ldots)\,, \quad \varphi_k(w; x) = y_k\,, \tag{8.3}$$

a mapping

$$\varphi : \mathcal{W} \times G \to \mathcal{G} = \{(y_0, y_1, \ldots) : y_k \in G \text{ for all } k\}\,. \tag{8.4}$$

Moreover, we consider a terminal set $\mathcal{T} \subset G$ and define

$$N_{\mathcal{T}}(w; x) = \min\{k : k \in \mathbb{N}_0\,,\, \varphi_k(w; x) \in \mathcal{T}\}\,. \tag{8.5}$$

The cost functional is given by

$$J(w; x) = \sum_{k=0}^{N-1} c(y_k, w_k) + c_{\mathcal{T}}(y_N)\,, \quad N = N_{\mathcal{T}}(w; x)\,, \tag{8.6}$$

if $N_{\mathcal{T}}(w; x) < +\infty$; if not, we set

$$J(w; x) = +\infty\,. \tag{8.7}$$

The function $c : G \times W \to \mathbb{R}$ denotes the running costs, the function $c_{\mathcal{T}} : G \to \mathbb{R}$ the terminal cost. For a given initial value $x \in G$ we want to find an optimal w_*, that is, a $w_* \in \mathcal{W}$ such that

$$J(w_*; x) = \min_{w \in \mathcal{W}} J(w; x)\,. \tag{8.8}$$

Definition 8.2 (Optimal value function)
The function $V : G \to [-\infty, +\infty]$ defined by

$$V(x) = \inf_{w \in \mathcal{W}} J(w; x) \tag{8.9}$$

*is called the **optimal value function** for Problem 8.1. A control $w_* \in \mathcal{W}$ is called optimal for the initial value $x \in G$ if*

$$V(x) = J(w_*; x)\,. \tag{8.10}$$

□

Definition 8.3 (Feedback control)
Let a discrete evolution $g : G \times W \to G$ be given. Any mapping $\omega : G \to W$ is called a ***feedback control****. We define the sequence of states belonging to ω by*

$$y_{k+1} = g(y_k, \omega(y_k)) \,, \quad y_0 = x \,. \tag{8.11}$$

A feedback control ω_* is called ***optimal*** *if (8.10) holds for*

$$w_* = (\omega_*(x), \omega_*(y_1), \ldots) \,, \quad \textit{for all } x \in G \,. \tag{8.12}$$

□

Proposition 8.4 (Bellman principle, principle of dynamic programming)
The optimal value V of problem 8.1 satisfies

$$V(x) = \inf_{v \in W} \left[c(x, v) + V(g(x, v)) \right] \,, \quad \textit{if } x \notin \mathcal{T} \,, \tag{8.13}$$

and $V(x) = c_{\mathcal{T}}(x)$, if $x \in \mathcal{T}$. If ω_* is an optimal feedback control, then

$$V(x) = c(x, \omega_*(x)) + V(g(x, \omega_*(x))) = \inf_{v \in W} \left[c(x, v) + V(g(x, v)) \right] \,, \quad \textit{if } x \notin \mathcal{T} \,. \tag{8.14}$$

Proof: For $v \in W$ and $w \in \mathcal{W}$ we set

$$(v, w) = (v, w_0, w_1, \ldots) \,. \tag{8.15}$$

Then

$$\varphi((v, w); x) = (x, \varphi(w; g(v, x))) \,, \tag{8.16}$$

holds for all $v \in W$, $w \in \mathcal{W}$ and $x \in G$, thus

$$J((v, w); x) = c(x, v) + J(w; g(x, v)) \,. \tag{8.17}$$

The first assertion now follows from

$$\begin{aligned} V(x) &= \inf_{w \in \mathcal{W}} J(w; x) = \inf_{\substack{v \in W \\ w \in \mathcal{W}}} J((v, w); x) \\ &= \inf_{\substack{v \in W \\ w \in \mathcal{W}}} \left[c(x, v) + J(w; g(x, v)) \right] = \inf_{v \in W} \left[c(x, v) + \inf_{w \in \mathcal{W}} J(w; g(x, v)) \right] \\ &= \inf_{v \in W} \left[c(x, v) + V(g(x, v)) \right] \,. \end{aligned} \tag{8.18}$$

If ω_* is an optimal feedback control, then for every $x \in G$ and the corresponding optimal control $w_* = (\omega_*(x), \tilde{w}_*)$ we obtain

$$\begin{aligned} V(x) = J(w_*, x) &= c(x, \omega_*(x)) + J(\tilde{w}_*, g(x, \omega_*(x))) \geq c(x, \omega_*(x)) + V(g(x, \omega_*(x))) \\ &\geq V(x) \,, \end{aligned} \tag{8.19}$$

by definition of $V(g(x, \omega_*(x)))$ and because of (8.13). □

Proposition 8.5 *Let ω_* be a feedback control which satisfies*

$$V(x) = c(x, \omega_*(x)) + V(g(x, \omega_*(x)))\,, \quad \text{if } x \notin \mathcal{T}\,. \tag{8.20}$$

Assume that $V(x) > -\infty$ for all $x \in G$ and

$$J(w_*; x) < +\infty\,, \quad \text{if } V(x) < +\infty\,. \tag{8.21}$$

Then ω_* is an optimal feedback control.

Proof: We have to show that

$$V(x) = J(w_*; x) \tag{8.22}$$

holds for all $x \in G$ with $V(x) < \infty$. We define

$$G_k = \{x : x \in G\,,\ N_{\mathcal{T}}(w_*; x) = k\}\,, \tag{8.23}$$

and use induction to prove that (8.22) holds in G_k. For $x \in G_0$ we have $V(x) = c_{\mathcal{T}}(x) = J(w_*; x)$. Let $x \in G_{k+1}$, then $g(x, \omega_*(x)) \in G_k$. Now the induction hypothesis implies, setting $w_* = (\omega_*(x), \tilde{w}_*)$,

$$V(x) = c(x, \omega_*(x)) + V(g(x, \omega_*(x))) = c(x, \omega_*(x)) + J(\tilde{w}_*; g(x, \omega_*(x))) = J(w_*; x)\,. \tag{8.24}$$

□

Problem 8.6 (Continuous control problem)
Minimize

$$J(w; x, t) = \int_t^T L(s, y(s), w(s))\, ds + L_1(y(T))\,. \tag{8.25}$$

Here, $y : [t, T] \to \mathbb{R}^n$ solves the initial value problem

$$\dot{y} = f(s, y, w(s))\,, \quad y(t) = x\,, \tag{8.26}$$

the final time $T \in \mathbb{R}$ is given, and the controls belong to the set

$$w \in \mathcal{W}_t = \{w |\, w \in L^\infty(t, T),\ w(s) \in W \text{ a.e.}\}\,, \tag{8.27}$$

with a given measurable set $W \subset \mathbb{R}^m$. □

Problem 8.6 represents a family $P_{x,t}$ of optimization problems which are parametrized by the initial value $(x, t) \in \mathbb{R}^n \times [t_0, T]$, $t_0 < T$ is given.

Assumption 8.7 *We assume that the initial value problem (8.26) has a unique solution $y : [t, T] \to \mathbb{R}^n$ for every $w \in \mathcal{W}_t$ and every $(x, t) \in \mathbb{R}^n \times [t_0, T]$. We also assume that $L : [t_0, T] \times \mathbb{R}^n \times W \to \mathbb{R}$ is continuous and that*

$$s \mapsto L(s, y(s), w(s))$$

is integrable on $[t, T]$ for all initial values (x, t) and all $w \in \mathcal{W}_t$ with corresponding solutions y of (8.26).

Definition 8.8 (Optimal value function)
We define the optimal value function $V : \mathbb{R}^n \times [t_0, T] \to [-\infty, +\infty]$ *by*

$$V(x,t) = \inf_{w \in \mathcal{W}_t} J(w; x, t)\,. \tag{8.28}$$

A control $w_* \in \mathcal{W}_t$ *is called optimal if*

$$V(x,t) = J(w_*; x, t)\,. \tag{8.29}$$

□

Proposition 8.9 (Bellman principle, continuous case)
Let assumption 8.7 hold. Then the optimal value function V *of problem 8.6 satisfies, for all* $(x,t) \in \mathbb{R}^n \times [t_0, T]$,

$$V(x,t) = \inf_{w \in \mathcal{W}_t} \left[\int_t^\tau L(s, y(s), w(s))\, ds + V(y(\tau), \tau) \right], \quad \forall\, \tau \in [t, T]\,, \tag{8.30}$$

where y *solves (8.26), as well as*

$$V(x,T) = L_1(x)\,, \quad x \in \mathbb{R}^n\,. \tag{8.31}$$

If $w_* \in \mathcal{W}_t$ *is optimal, then*

$$V(x,t) = \int_t^\tau L(s, y_*(s), w_*(s))\, ds + V(y_*(\tau), \tau)\,, \quad \forall\, \tau \in [t, T]\,. \tag{8.32}$$

Proof: We first prove that "$\leq$" holds in (8.30). Let $w \in \mathcal{W}_t$, $\tau \in [t, T]$, let $\tilde{w} \in \mathcal{W}_\tau$ be arbitrary. We define

$$\overline{w}(s) = \begin{cases} w(s)\,, & s \in [t, \tau)\,, \\ \tilde{w}(s)\,, & s \in [\tau, T]\,. \end{cases}$$

We have $\overline{w} \in \mathcal{W}_t$, and the corresponding state function $\overline{y}$ is given by

$$\overline{y}(s) = \begin{cases} y(s)\,, & s \in [t, \tau)\,, \\ \tilde{y}(s)\,, & s \in [\tau, T]\,. \end{cases}$$

From the definition of V we obtain

$$\begin{aligned} V(x,t) &\leq \int_t^T L(s, \overline{y}(s), \overline{w}(s))\, ds + L_1(\overline{y}(T)) \\ &= \int_t^\tau L(s, y(s), w(s))\, ds + J(\tilde{w}; \tau, y(\tau))\,. \end{aligned} \tag{8.33}$$

Passing to the infimum w.r.t. $\tilde{w} \in \mathcal{W}_\tau$ and $w \in \mathcal{W}_t$ yields "$\leq$".
Let now $\delta > 0$ be arbitrary, let $w \in \mathcal{W}_t$ mit $J(w; x, t) \leq V(x,t) + \delta$. Then

$$\begin{aligned} V(x,t) + \delta &\geq J(w; x, t) \\ &= \int_t^\tau L(s, y(s), w(s))\, ds + \int_\tau^T L(s, y(s), w(s))\, ds + L_1(y(T)) \\ &= \int_t^\tau L(s, y(s), w(s))\, ds + J(w; y(\tau), \tau) \\ &\geq \int_t^\tau L(s, y(s), w(s))\, ds + V(y(\tau), \tau) \\ &\geq V(x,t)\,, \end{aligned} \tag{8.34}$$

the last inequality follows from the already proven inequality "$\leq$".
If $w = w_*$ is optimal, (8.34) holds with $\delta = 0$, therefore (8.32) follows. □

Proposition 8.10 (Hamilton-Jacobi-Bellman equation)
Let $(x,t) \in \mathbb{R}^n \times (t_0, T)$, let assumption 8.7 hold. If the optimal value function V is differentiable in the point (x,t), we have

$$\partial_t V(x,t) + \langle \nabla V(x,t), f(t,x,v) \rangle + L(t,x,v) \geq 0\,, \quad \text{for all } v \in W. \tag{8.35}$$

If w_* is an optimal control, V is differentiable in the point $(y_*(t),t)$ and w_* is continuous in t, we have

$$\partial_t V(x,t) + \min_{v \in W} [\langle \nabla V(x,t), f(t,x,v) \rangle + L(t,x,v)] = 0\,, \tag{8.36}$$

and the minimum is attained at the value $v = w_*(t)$.

Proof: Let $v \in W$ be arbitrary, let $\delta < T - t$. We choose a $w \in \mathcal{W}_t$ satisfying $w|[t, t+\delta] = v$. Let y be the corresponding state function. From (8.30) it follows for all $h \in (0, \delta)$ that

$$\frac{V(y(t+h), t+h) - V(x,t)}{h} \geq -\frac{1}{h} \int_t^{t+h} L(s, y(s), v)\, ds\,.$$

Passing to the limit $h \downarrow 0$ yields (8.35). In the same way it follows from (8.32) that

$$\frac{V(y_*(t+h), t+h) - V(y_*(t),t)}{h} = -\frac{1}{h} \int_t^{t+h} L(s, y_*(s), w_*(s))\, ds\,.$$

Letting $h \downarrow 0$ shows that the minimum value 0 is attained at $v = w_*(t)$. □

Proposition 8.10 shows that the optimal value function V solves, at all points (x,t) where it is differentiable, a so-called Hamilton-Jacobi equation

$$\partial_t u + H(x,t,\nabla u) = 0\,, \tag{8.37}$$

where in our case

$$H(x,t,p) = \min_{v \in W} [\langle p, f(t,x,v) \rangle + L(t,x,v)]\,. \tag{8.38}$$

When H has the special form (8.38), (8.37) is called the **Hamilton-Jacobi-Bellman equation**, in short **HJB equation**.

Proposition 8.11 (Sufficient optimality conditions)
Let 8.7 hold, let $u : \mathbb{R}^n \times [t_0, T] \to \mathbb{R}$ be continuously differentiable with

$$u(x,T) = L_1(x)\,, \quad x \in \mathbb{R}^n\,, \tag{8.39}$$

$$\partial_t u(x,t) + \inf_{v \in W} [\langle \nabla u(x,t), f(t,x,v) \rangle + L(t,x,v)] = 0\,, \quad \forall\, (x,t) \in \mathbb{R}^n \times [t_0, T)\,. \tag{8.40}$$

Then

$$u(x,t) \leq V(x,t)\,, \quad \forall\, (x,t) \in \mathbb{R}^n \times [t_0, T)\,. \tag{8.41}$$

If moreover $w_* \in \mathcal{W}_t$ is a control satisfying

$$\partial_t u(y_*(s), s) + \langle \nabla u(y_*(s), s), f(s, y_*(s), w_*(s)) \rangle + L(s, y_*(s), w_*(s)) = 0\,, \tag{8.42}$$

for almost all $s \in [t_0, T]$, then w_* is an optimal control for $P_{(t,x)}$, and

$$u(x,t) = V(x,t)\,. \tag{8.43}$$

Proof: Let $(x,t) \in \mathbb{R}^n \times [t_0, T)$, $w \in \mathcal{W}_t$ be arbitrary. Then

$$\begin{aligned} J(w;x,t) &= \int_t^T L(\tau, y(\tau), w(\tau))\, d\tau + L_1(y(T)) \\ &= \int_t^T L(\tau, y(\tau), w(\tau))\, d\tau + u(x,t) + \int_t^T \frac{\mathrm{d}}{\mathrm{d}\tau} u(y(\tau),\tau)\, d\tau \\ &= \int_t^T L(\tau, y(\tau), w(\tau))\, d\tau + u(x,t) + \\ &\quad + \int_t^T \partial_t u(y(\tau),\tau) + \langle \nabla u(y(\tau),\tau), f(\tau, y(\tau), w(\tau))\rangle \; d\tau \\ &\geq u(x,t)\,, \end{aligned} \tag{8.44}$$

therefore (8.41) follows. If w_* yields the minimum in (8.40), the same computation shows that

$$J(w_*; x, t) = u(x,t) \leq V(x,t)\,.$$

Therefore w_* is optimal and $u(x,t) = V(x,t)$. □

Remark 8.12 (Construction of optimal feedback controls)
If the optimal value function is continuously differentiable, one can construct from Proposition 8.11 the optimal value function as well as the optimal control in feedback form in the following manner.

1. Determine $\omega(x,t,p)$, $p \in \mathbb{R}^n$ such that the minimum w.r.t. $v \in W$ of

   $$\langle p, f(t,x,v)\rangle + L(t,x,v)$$

   is attained at the value $v = \omega(x,t,p)$.

2. Solve the partial differential equation

   $$\partial_t u + \langle \nabla u, f(t,x,\omega(x,t,\nabla u))\rangle + L(t,x,\omega(x,t,\nabla u)) = 0\,, \tag{8.45}$$

   with the boundary condition

   $$u(x,T) = L_1(x)\,, \quad x \in \mathbb{R}^n\,. \tag{8.46}$$

3. Check whether the solution satisfies the smoothness requirements of Proposition 8.11.

4. Compute y_* as the solution of

   $$y' = f(s, y, \omega(y,s,\nabla u(y,s))\,, \quad y(t) = x\,, \tag{8.47}$$

   and the optimal control w_* from

   $$w_*(s) = \omega(y_*(s), s, \nabla u(y_*(s), s))\,. \tag{8.48}$$

Remark 8.13 (Unfortunately ...)

- Only in very special cases the optimal value function is globally continuously differentiable. An example is given by the linear-quadratic problem.
- The construction method 8.12 may work in cases where the optimal value function is piecewise continuously differentiable and the set on which ∇V is discontinuous consists of smooth surfaces. It may then be possible to apply the construction method between such surfaces successively. But those surfaces also have to be determined.
- Until around 1980 this was the only method to solve the HJB equation. The situation changed when the concept of viscosity solutions was invented.

Let us consider the linear-quadratic problem.

Problem 8.14 (Linear-quadratic problem)
Minimize

$$J(w;x,t) = \int_t^T x(s)^T M(s)x(s) + w(s)^T N(s)w(s)\,ds + x(T)^T Dx(T)\,, \tag{8.49}$$

subject to the constraints

$$x' = A(s)x + B(s)w\,, \quad x(t) = x\,, \tag{8.50}$$

with given terminal time T and without control constraint, that is, $W = \mathbb{R}^m$.

□

Here we use the letter x (instead of y) for the state function and accept the notation $x(t) = x$ for the initial condition.

Assumption 8.15 *Let A, B, M, N be continuous matrix-valued functions of s of suitable dimension, let $D, M(s), N(s)$ be symmetric for all s, let $D, M(s)$ be positive semidefinite and $N(s)$ positive definite for all s.* □

Ansatz 8.16
We make the ansatz

$$u(x,t) = x^T Q(t)x = \langle x, Q(t)x\rangle\ , \tag{8.51}$$

where $Q(t)$ is a symmetric matrix in $\mathbb{R}^{(n,n)}$ for every t, and $t \mapsto Q(t)$ is continuously differentiable. In order to compute the optimal control, for every $(x,t) \in \mathbb{R}^n \times \mathbb{R}$, $t < T$ we have to minimize the function

$$g(v) = \langle \nabla u(x,t), f(t,x,v)\rangle + L(t,x,v) \tag{8.52}$$

over $v \in W = \mathbb{R}^m$. Inserting f and L yields (we omit the argument t)

$$\begin{aligned} g(v) &= \langle x, Q(Ax+Bv)\rangle + \langle Ax+Bv, Qx\rangle + x^T Mx + v^T Nv \\ &= 2x^T Q(Ax+Bv) + x^T Mx + v^T Nv\,, \end{aligned} \tag{8.53}$$

since we have assumed Q to be symmetric. Because N is positive definite, g is strictly convex. Every zero of ∇g is then a global minimizer of g. We have

$$\nabla g(v) = 2x^T QB + 2v^T N\,, \tag{8.54}$$

therefore

$$\nabla g(v) = 0 \quad \Leftrightarrow \quad 2x^T QB = -2v^T N \quad \Leftrightarrow \quad v = -N^{-1}B^T Qx\,. \tag{8.55}$$

(N^{-1} exists since N is positive definite.) We thus have shown: It follows from the (8.51) with Q symmetric that

$$\omega(x,t) = -N(t)^{-1}B(t)^T Q(t)x \tag{8.56}$$

is the unique solution of

$$\min_{v\in\mathbb{R}^m} \left[\langle \nabla u(x,t), f(t,x,v)\rangle + L(t,x,v)\right]\,. \tag{8.57}$$

In order to determine $Q(t)$ we consider the HJB equation

$$\partial_t u(x,t) + \min_{v\in\mathbb{R}^m} \left[\langle \nabla u, f(t,x,v)\rangle + L(t,x,v)\right] = 0\,. \tag{8.58}$$

Choosing $v = \omega(x,t)$ from (8.56) and u from (8.51) we obtain

$$\begin{aligned}
0 &= \partial_t u(x,t) + \langle \nabla u, f(t,x,\omega(x,t))\rangle + L(t,x,\omega(x,t)) \\
&= \langle x, Q'x\rangle + \langle x, Q(Ax+B\omega)\rangle + \langle Ax+B\omega, Qx\rangle + x^T Mx + \omega^T N\omega \\
&= x^T Q'x + x^T(QA + A^T Q + M)x + x^T QB\omega + \omega^T B^T Qx + \omega^T N\omega \\
&= x^T(Q' + QA + A^T Q + M)x + x^T QB(-N^{-1})B^T Qx \\
&\quad - x^T QBN^{-1}B^T Qx + x^T QBN^{-1}NN^{-1}B^T Qx \\
&= x^T \underbrace{(Q' + QA + A^T Q + M - QBN^{-1}B^T Q)}_{=:P} x\,.
\end{aligned} \tag{8.59}$$

If we can find a symmetric $Q(t)$ such that $P(t) = 0$ for all t and $Q(T) = D$, then the function u from (8.51) satisfies the sufficient conditions in Proposition 8.11, and we have computed both the optimal value function and an optimal control in feedback form. □

Proposition 8.17 (Solution of the linear-quadratic problem)
Let the assumptions 8.15 hold. Then the backward initial value problem for the so-called matrix Riccati differential equation

$$Q' = -QA - A^T Q - M + QBN^{-1}B^T Q^T\,, \quad Q(T) = D\,, \tag{8.60}$$

has a unique solution $Q : (-\infty, T] \to \mathbb{R}^{(n,n)}$, and $Q(t)$ is symmetric for all $t \le T$. The linear-quadratic problem 8.14 has a unique solution w_* for all $(x,t) \in \mathbb{R}^n \times \mathbb{R}$ with $t \le T$. This solution is obtained in the following manner.

1. *Solve (8.60).*

2. *Set*

$$\omega(x,t) = -N(t)^{-1}B(t)^T Q(t)x\,, \tag{8.61}$$

and determine the unique solution x_* of the initial value problem

$$x' = A(s)x + B(s)\omega(x,s)\,, \quad x(t) = x\,, \tag{8.62}$$

with $\omega(x,t)$ from (8.61).

3. *Set*

$$w_*(s) = \omega(x_*(s),s)\,, \quad s \in [t,T]\,. \tag{8.63}$$

The optimal value function of Problem 8.14 is given by

$$V(x,t) = x^T Q(t)x\,. \tag{8.64}$$

Proof: Let $Q : [t_0,T] \to \mathbb{R}^{(n,n)}$ be a solution of (8.60) for some $t_0 < T$. Then $R = Q - Q^T$ solves the backward initial value problem

$$R' = -RA - A^T R\,, \quad R(T) = 0$$

on $[t_0,T]$, therefore $R = 0$ and thus Q is symmetric. We set

$$u(x,t) = x^T Q(t)x\,. \tag{8.65}$$

Then, according to the construction in 8.16, the assumptions of Proposition 8.11 are satisfied.
It remains to prove that the local solution of (8.60) can be extended to $(-\infty,T]$. Let $(t_-,T]$ be the maximal existence interval of the solution. Since $D, M(t)$ and $N(t)$ are positive semidefinite, the cost function J of the original problem is nonnegative for all controls, and therefore

$$0 \le V(x,t) = u(x,t) = x^T Q(t)x \tag{8.66}$$

holds for all $t \in (t_-,t_1]$ and all $x \in \mathbb{R}^n$. Since the control $\tilde{w} = 0$ is admissible, for all $t \in (t_-,t_1]$ the state $\tilde{x}$ belonging to the initial value (x,t) satisfies (here Φ denotes the transition matrix of the system $x' = A(s)x$)

$$\tilde{x}(s) = \Phi(s,t)x\,,$$

and moreover

$$\begin{aligned} 0 \le x^T Q(t)x \le J(\tilde{w};x,t) &= \int_t^T \tilde{x}(s)^T M(s)\tilde{x}(s)\,ds + \tilde{x}(T)^T D\tilde{x}(T) \\ &\le |x|^2 \int_t^T \|\Phi(s,t)\|_2^2 \|M(s)\|_2\,ds + |x|^2\|D\|_2\|\Phi(T,t)\|_2^2 \\ &= c(t)|x|^2\,, \end{aligned} \tag{8.67}$$

where $c : (-\infty,T]$ denotes the function which arises from the foregoing line; it is continuous. From

$$0 \le x^T Q(t)x \le c(t)|x|^2\,, \quad Q \text{ symmetric}, \tag{8.68}$$

it follows that

$$\|Q(t)\|_2 \leq c(t)\,.$$

Now, if $t_- > -\infty$, then $Q(t)$ is bounded on $(t_-, T]$. But then the solution of (8.60) can be extended to the left of t_-, a contradiction since t_- is maximal. □

The feedback control ω from (8.61) is not a pure state feedback, because the time t arises explicitly. This is also the case in the autonomous LQ problem (with constant matrices A, B, M, N) since Q always depends on t.